Peter R. Bergethon

The Physical Basis of Biochemistry

The Foundations of Molecular Biophysics

Second Edition

Springer

The Physical Basis of Biochemistry

Peter R. Bergethon

The Physical Basis of Biochemistry

The Foundations of Molecular Biophysics

Second Edition

Chapters 27 and 28 written with Kristin E. Bergethon

Springer

Peter R. Bergethon
Departments of Anatomy & Neurobiology
and Biochemistry
Boston University School of Medicine
715 Albany Street
Boston, Massachusetts 02118-2526
USA
prberget@bu.edu

ISBN 978-1-4419-6323-9 e-ISBN 978-1-4419-6324-6
DOI 10.1007/978-1-4419-6324-6
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2010929616

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface to the Second Edition

In its first edition, the Physical Basis of Biochemistry endeavored to connect the foundation principles of biology to the essential processes in physics and chemistry. The core purpose of the new edition remains that described in the first edition:

- *The Physical Basis of Biochemistry* is an introduction to the philosophy and practice of an interdisciplinary field in which biological systems are explored using the quantitative perspective of the physical scientist. I have three primary objectives in this volume: one, to provide a *unifying* picture of the interdisciplinary threads from which the tapestry of biophysical studies are woven; two, to provide an insight into the power of the *modeling* approach to scientific investigation; and three, to communicate a sense of *excitement* for the activity and wholesome argument that characterizes this field of study.
- The students in this course are likely to be a cross section of quantitatively oriented biologists as well as aspiring chemists, physicists, and engineers with an interest in biological systems. This leads to a mixing of interests and disciplines brought to both the classroom and the laboratory. A fundamental assumption of a physical approach to biology and especially to biophysical chemistry is that we can gain understanding of a biological phenomenon by describing and understanding its physical nature. Thus the tools of the physical scientist become available to the biologist in the exploration of the complexities of biological systems.

In the preface to the first edition of The Physical Basis of Biochemistry I wrote

> In recent years, the skills of the biophysical chemist have emerged as prized commodities in both academic and industrial circles. The need for these skills will intensify as the practice in biotechnology and bioindustry moves from gene identification to the manufacture of gene products that will need both proper structure and function to be useful.

Just 6 years after that preface was written, the human genome was decoded and in the last decade genomics has given way to proteomics and the apparent rediscovery of the biological system. There have been several “great surprises” such as the complexity of the protein-folding question, the massive diversity of form and function that arises from a much smaller number of genes than expected, and an explosion of biomathematical needs and expectations as the field of bioinformatics has become critical to application of these data to biological questions in fields as diverse as evolution, anthropology, linguistics, neurosciences, and biomedicine.

The idea that a fundamental understanding of the basic physical principles underlying chemical biological systems is vital remains the focus of this new edition. This new edition has been extensively reworked and reorganized to be more pedagogically friendly to the original goals as outlined above. The parallel construction of topics in the first edition was much more difficult to teach in classrooms than expected and a more traditional ordering has now been restored. There has been substantial new material added with respect to models at the simple molecular level—including van der Waals gases and virial treatments—which is tied to coverage of models of polymer thermodynamics. The methods for biophysical analysis have largely been reorganized and are now found in Part V.

The new edition is again partitioned into five parts:

Part I explores the central concept that science is a way of looking at the world. The role of scientific inquiry and its dependence on systems analysis and model making (the progression of inquiry) is once again emphasized with respect to building a background in biological content, an approach to systems science, and a review of probability and statistics.

Part II reviews the physical underpinnings of biophysical chemistry with an emphasis first on energy, work, and forces of biological importance. Then an introduction to quantum mechanics, chemical principles, and thermodynamics prepares the student with the tools necessary for constructing models in biological state space.

Part III uses the physical foundations developed in the first two parts to explore how models applicable to molecular biophysics are constructed. The overall system is that of aqueous biochemistry. The part starts with a consideration of the properties of water and then sequentially explores the interactions of water with the chemical components that make up biological systems. In the development of ion–solvent and ion–ion models, the Born, Kirkwood, and Debye–Hückel models instruct us on how these great physical scientists brought us to where we are today. There is a certain "classical" coverage of material but it is important for the modern student to see where the simplifications and abstractions originated in many of the ideas that have become our modern dogma. The part explores the interactions that lead to macromolecular (polymer) structure, the properties of the cell membrane, and the structure of the electrified regions near cells and colloidal surfaces.

Part IV takes the three-dimensional potential energy surface, which reflects the equilibrium state, and puts it into motion. This part explores the time-dependent actions of real-world processes. The major driving forces in biological systems are chemical and electrical gradients, and, diffusion and conduction are the focus of this part. It concludes with an examination of the electromechanical phenomena of electrokinetics and the kinetics of chemical and electrochemical systems of biological importance: enzymes and electron transfer in proteins.

Part V is a succinct discussion of the biophysical methods used to evaluate structure and function in biological systems. First the physics underlying the methods that use mechanical macroscopic properties to cause motion in a field such as centrifugation, electrophoresis, mass spectrometry, and chromatography are presented. Then the exploration of molecular structure with photons through spectroscopy and scattering techniques is presented. Finally the use of imaging techniques such as light, fluorescence, and atomic force microscopy is examined.

The *Appendices* continue to serve the purpose of presenting in detail some review information and certain topics that will be of interest to some readers, but might otherwise disrupt the flow of the textual story line.

The *question and problem* coverage in this volume has undergone some moderate expansion but a companion problems and solutions manual covering the material in this edition and much more extensively providing exercises to support learning these topics is currently being written and should be available concurrently with this text.

As always, projects like this never happen in a vacuum and I remain indebted to the help and support provided by colleagues and my family as this project now reaches into its third decade! This text has become an intergenerational family business. Like the first edition, in which I had the pleasure of working with my late father, Kaare Roald Bergethon, who contributed his scholarly and linquistic skills, in this edition I am pleased to have collaborated with my daughter, Kristin Elizabeth Bergethon, who has come into her own as a chemist and made substantial contributions to several of the chapters in Part 5. Unfortunately, this edition had to proceed without the encouragement of my colleague and friend Mary T. Walsh whose untimely death has left the community without a dedicated teacher and researcher. She is missed by all who knew her. In addition I would like to acknowledge the efforts to improve this edition hopefully will reflect well on those whose input has only helped to accomplish the goal. Any failures and errors that remain are entirely my responsibility.

Boston, Massachusetts
September 2009

Peter R. Bergethon

Contents

Part III Building a Model of Biomolecular Structure

Part IV Function and Action Biological State Space

Part I
Principles of Biophysical Inquiry

Chapter 1
Introduction: To the Student – First Edition

Fundamentally this book is about biology.

It is true that there are plenty of mathematical phrases and expressions, but this is because mathematics is the best language to use when we talk about many of these subjects. It is important to be comfortable with the languages of science, and mathematics is one of the most elegant. I have felt free to use mathematical development with many of the subjects, and there is a real biological method to this madness. The really interesting part of biology – and what makes a biologist a biologist – is that our attention is continually drawn past the details of the individual elements of a living system to the way the system acts as a whole.

The biologist, having discovered the cell, imagines the organization of cells into organs, organs into organisms, organisms into communities, communities into ecosystems, and ecosystems into evolutionary processes and trends. The philosophical pathway here is the *system* of individual elements. The biologist is a system scientist, and a good biologist is rigorous about describing the elements of the system. Without a precise way of talking about the details of the system, there is no adequate way to describe the system. If you think there is no rigor or dense challenging language in systematic or evolutionary biology read the detailed descriptions of Darwin’s finches’ beaks or the Latin used to describe biological species.

For the biologist who works in biophysics, biochemistry, or cell and molecular biology, the key is still the system. The details are now physical forces and properties of molecular and atomic structures, but the focus for the biologist is how these elements interact together. These interactions allow arrangement of ions, carbon skeletons, nitrogen, oxygen, sulfur, and other atoms into biologically active structures. The interactions

- determine how a protein folds,
- form membranes,
- determine how membranes and enzymes extract energy, and
- organize membranes and enzymes to assemble new structural elements.

These same interactions also determine the structure of information needed by the biological system. They operate in the communication

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_1, © Springer Science+Business Media, LLC 2010

- between organelles (such as second messenger systems),
- between cells and organs (nerve impulses), and
- over generations (DNA).

For the biologist interested in the organization and control mechanisms at the cellular and subcellular level, the language of the system is not Latin but physics and the formalisms of mathematics.

So here lies the essence of my passion for the subject: if we are interested in defining and understanding *biological systems* at a molecular level we must speak the language and understand the principles applicable at that level. The trick is to know which details are necessary and when. *This is the art of abstraction.* Physical details are often a great gift to the biologist, for it is the biologist, both by training and by inclination, who appreciates the whole picture of living systems. Remember Darwin and his finches: it was the details of the beaks and the types of seeds each finch had become adapted to eat that lead him to the *biological* principle of evolution. To formulate evolution he had to understand levers, torque, wedges, and a variety of physical principles that let him see past the physics and to deduce the intrinsic order of the biological system.

So it remains in our time. We must know the physical and chemical principles well enough that we can examine biological systems and see beyond to the biological organization making up the system. The curriculum for our study of living systems is not defined by the wishes of the faculty or the student: it is defined by the natural system we chose to investigate. Our free will starts and ends with this choice. Once we decide to study biological systems at the molecular level, we must accept without shirking that we are committed to a study of complex systems, thermodynamics, concentrated solutions, surface dominated non-homogenous chemistry, organic and organometallic chemistry, electrostatics, electrochemistry, and mechanics (both quantum and classical) just to name a few. Aspects of these subjects are challenging, but a practical knowledge of them is an attainable goal. The journey to acquire them is a wondrous one. Our passport is the art of the abstraction that allows complex subjects to be reduced to fundamental principles, and biophysical chemistry brings together the building blocks essential to biological investigation.

Chapter 2
Philosophy and Practice of Biophysical Study

Contents

2.1 What Is Biophysical Chemistry and Why Study It?

As a field, biophysical chemistry is an interdisciplinary area of study in which biological systems are regarded with the somewhat quantitative and concrete eye of the physical scientist. In using the intellectual paradigm of biophysical chemistry, we attempt to understand a biological phenomenon by carefully describing the essentials of its physical nature. This gives us the advantage of using the tools of the physical scientist to explore the complexities of biological systems. These tools are essentially the language and formalisms of mathematics, physics,

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_2, © Springer Science+Business Media, LLC 2010

and chemistry. The underlying philosophical foundation of biophysical chemistry is that application of the principles of these fields to biological systems will lead to meaningful and useful knowledge. Although it is impossible to advance understanding of biological problems using a biophysical paradigm without being knowledgeable about the underlying physical and chemical principles, when teaching and presenting these fundamentals, both the instructor and the student tend to get lost in the physical details and forget the overall purpose for the investigation. In this volume we will endeavor to find the proper balance.

Recently a newer term, *molecular biophysics* is often found alongside the more traditional name, biophysical chemistry. Is there a difference between these two topics? In this text, we will recognize the distinction but regard them as not being very different. In practice (such as the naming of courses of study at a University), these two topics are often differentiated but from our philosophical worldview, it is far more cogent to view them as the same. Operationally they might be differentiated by the property that biophysical chemistry is “intellectually” willing to embrace topics at the phenomenological level and proceed with physical analysis even when a molecular basis and understanding is not yet known. Molecular biophysics takes as its starting point that some knowledge of the molecule is given and then applies physical analysis. Sometimes in biophysical chemistry, progress is made and the behavior of the system under study is elucidated even when molecular detail is unavailable. There is the expectation this approach will then lead to a more detailed molecular-explanation level for the phenomenon in due course. The biophysical chemist will therefore be able to measure and quantitatively model complex phenomena such as the nerve action potential or protein folding in physical terms before understanding the physical structure of the molecules that give rise to the behavior. The molecular biophysicist would be more inclined to focus first on a detailed description and understanding of the molecule before attempting to understand its integration and response to its physical environment. The researcher whose efforts use physical methods such as x-ray crystallography or NMR spectroscopy to define the structure of a macromolecule such as a sodium channel would be inclined to identify oneself as a molecular biophysicist (or structural biologist). Alternatively, the scientist seeking the relationship of how mechanical traction alters the electrical conductivity of a lipid membrane in aqueous electrolyte solution of NaCl might call oneself a biophysical chemist. We will take the position that these are two ends of the same field of inquiry – both cases of the application of the fundamental study of physics, which is the search for the fundamental set of rules and actors in such a fashion that general principles can be inferred that are explanatory and predictive when applied to often dizzyingly complex and diverse biological systems. We will generally refer to the overall field as biophysical chemistry but will use the terms interchangeably guided by the context – always recognizing that the underlying scientific viewpoint is the same.

2.2 Science Is Not Content but a Unique Method of Discovery

Biophysical chemistry is a branch of *modern scientific inquiry* that uses a physical perspective in its application of the scientific method. What we mean by the *scientific method* needs precise definition because it differentiates modern science from other methods of exploration and explanation. Science is defined as a human endeavor that explores the natural world. By no means is science the only method for exploring the natural world but it *is* distinct from other means. What makes modern scientific inquiry distinct is an insistence for *validation* of observations and relationships with a *skeptical empiricism* that requires both rational, *logical reasoning* and *evidence* that is generated by carefully done *experimentation*. Content information (often casually called “the science” by both lay and scientific workers) is often considered to be the essence of a scientific discipline. This is an error. Content information is not science but is the product of scientific inquiry and its methods properly applied. Ideally content information that is discovered and validated by the methods of science should be identified with the adjective “scientific.” This specifically tags the subsequent information as resulting from an operation of scientific inquiry: the scientific method that has been applied to an object or event of interest. The process of scientific inquiry always starts with observations of some aspect of the natural world. The first step of scientific observation is characterized by systematically organizing these analyses into a description of the system being observed. This systematic organization characterizes each *modern*, *proto-*, and *Aristotelian* scientific inquiry. We will name this overall group *holo-scientific* in our following discussions. Because biophysical chemistry is a product of modern scientific inquiry, when the unmodified term is used, the reader should assume reference is being made to *modern* scientific inquiry.

It is useful in holo-scientific analysis to take the view that everything can be treated as a “system.” A system is described using the following definitions:

- A *system* is a set of elements (e.g., parts, events, components, objects both physical and metaphorical) that are connected and that form a complex whole.
- The *properties* of systems are typically emergent from the overall operation of the system.
- *Observables* are properties that are measurable.
- A *description of a system* includes a notation of its
  - *Elements,*
  - *Relationship rules* defining how the elements interact,
  - *Context or background space* in which the elements and rules are found and operate,
  - *State*, which is defined by a set of observable properties that are measured together. *Patterns* of observables are typically used to define the state of a system. The identification of these patterns to define the “state” of biological systems is the focus of the field of “systems biology,”
  - *Equations of state* that define how the observables are related,
  - Typically the elements comprising a system are themselves further describable as systems (i.e., they are *subsystems)*. System properties are usually distinct from the individual characteristics (or properties) of the subsystem components of the system.

2.3 The Progression of Inquiry Guides the Scientific Modeling Process

Models are partial descriptions (or abstractions) of systems of interest. A "good" model is an abstraction at an appropriate level of detail that "accurately" represents the reality of the system of interest. Staged model making is the central process of the method of modern science. The practice of a cycle of critical analysis, hypothesis creation and skeptical empiricism to confirm hypotheses, and observation is named "*the progression of inquiry*." Thus a description of the dynamic of modern science is based on three linked stages of model making. These are the following:

- Formation of *descriptive models* that represent observations (of systems of interest).
- Generation of *explanatory models* that embed hypothetical linkages of causality (relating how the system of interest works).
- Creation of *experimental models* that allow empirical testing of the hypothesized relationships of the explanatory and the epistemological nature of the descriptive models. This empirical testing validates or falsifies the models by comparison of predictions derived from theory (hypotheses) to the measured experimental data.

Many models employ the abstraction that the system under study is static or unchanging over time. Many systems are, in fact, unchanging or static in the time frame of interest so this can be an important and reasonable abstraction. However this abstraction is unreasonable when some aspect of the system as described above changes with time. When changes in system elements, rules of interaction, context, emergent properties, states, or equations of state over time are required for accurate description and explanation of a system under study, dynamic modeling is required. Dynamic modeling explicitly follows changes in the system over time and in many cases a single timescale is inadequate to the task. If necessary, multiple timescales must be incorporated in the model. The mathematical tractability of such models is inversely proportional to the dynamical complexity. Sometimes the dynamical model can be shown to arrive at a dynamic equilibrium and can thus be treated as if it were a static model unchanging over time.

At this point it is adequate to understand that the process of modeling identifies and abstracts certain system-descriptive details of a real system under study and maps them to a formal system that is the model. It is correct to conceptualize the process of modeling as an *operator* thus indicating a *mathematical operation* on a real system to produce the formal model. When we wish to invoke a systems modeling operation that transforms an observation into a description of system as given above it can be designated it with the operator symbol Υ. We write the modeling operator:

$$\Upsilon(x) \tag{2.1}$$

This says find the “system of x.” Thus, if we wish to describe the phenomenon of muscular action (which we will do later in the chapter) we could write this as $\Upsilon(\text{action}_{\text{muscle}})$. In Chapter 4 we will extend our treatment of model making and its discipline, systems science.

With the operation of making a systems description defined we can further recognize that the progression of inquiry or scientific method is also an operator, ΠE. The transformation of a descriptive model (which is the output of $\Upsilon(x)$) into an experimentally validated causal model is an iterative process that is supervised by the progression of inquiry and can be written as

$$\Pi E\,[\Upsilon(x)] \tag{2.2}$$

Thus we define the scientific method as a paradigm for inquiring into the natural world such that

(1) observations (x) are organized using systems analysis into descriptive models, $\Upsilon(x)$;
(2) testable hypotheses that may be either correlational or causal in nature are proposed as connecting the elements and relational rules that characterize the descriptive models. This leads the transform of the descriptive model into an explanatory model of linked hypotheses (a theoretical model);
(3) predictions of values expected to be taken by a dependent variable when an independent variable is set, result from phrasing the hypotheses as relationships linking these variables;
(4) a scientific experiment which is a formal mapping of the theoretical model is done in which only the independent variable(s) is altered. The design of the scientific experimental model is a large part of the day-to-day practice of scientific investigation;
(5) experimental evidence supporting validation of the tested hypothesis is established when the value of the predicted observable is consistent with measured experimental value.

The terminology, experimental methodologies, theoretical approaches, and history of a scientific discipline may vary between fields but this fundamental process of discovery and validation is a general and shared operation.

2.4 A Brief History of Human Methods of Inquiry Reveals Important Aspects of the Scientific Method

Arguably the scientific method (modern science) is one of the most successful discoveries in over 300,000 years of human endeavor. In the half millennium since its development and application, human civilization has seen an overwhelming evidence of the power of modern science (the process) and of scientific knowledge to alter and mostly improve the human condition. Advancing knowledge of the natural world in which we live (the prime objective of holo-scientific study) continues on an exponential rise because of the modern method. How did this remarkable process of progress come to pass? Since the whole of science is a human endeavor that explores the natural world, it is the human brain with its strengths and weaknesses that does the exploring. A sense of our modern view of science can be gained by following a brief history of holo-scientific thought.

All (holo) science starts with observations of the patterns of the natural world. Archeological evidence from cave paintings and the notching of bone and reindeer horns suggests that pre-historic humans were extremely careful in their recording of seasonal and temporal patterns. Such knowledge is acquired by simple observation. A hunter/gather society depends on this level of inquiry to know where and when animals gather, feed, obtain water, or sleep; to know when and where berries, shrubs, and flowers are located that will bear fruit; and to know the patterns of weather, drought, and flood so that migration ensures the survival of the society. An agricultural society also depends on this basic knowledge to know when to plant, when to reap, and when to gather and store food.

However, observation alone is not “modern” science; simple observation leads to “proto-science.” Proto-science accepts observations without question or verification. The human brain always tries to organize observations of its world into unambiguous models of cause and effect. This does not mean that the models are correct, only unambiguous. Models of cause and effect are built on a mode of explanation or a context. A mode of explanation establishes the way that cause-and-effect relationships explain the natural world. Historically, humans and their cultures have used several modes of explanation in their attempt to formulate the cause-and-effect models that give meaning to their observations of the natural world. Three important modes of explanation seen in the history of humans are as follows:

- Received knowledge
- Ways of knowing
- Modern skeptical-empiricism

Received knowledge. This mode of explanation attributes the cause for events to Gods, magic, and mystical powers. It leads to mythological and theological explanations for the events discovered in the natural world. Mystical attribution is usually based on “received knowledge” or knowledge that is passed “down” invariably by those who have previously “received these truths.” The observed evidence, which may be quite accurate and detailed, is interpreted in a God–demon–magic context. For much of human history, patterns of importance to hunter-gatherer and agricultural civilizations have been used to ensure survival of the society. For most of human experience, the causality relations were attributed to supernatural Gods and magical occurrences. Thus, a “proto-science” based on observation with theological attribution has existed for most of humankind’s history.

Ways of knowing. The ancient Greeks made extremely careful observations about their natural world. They developed models that explained the observations according to strictly rational, logical deductions. The starting point for these deduced models was derived from “self-evident” truths. These models of thought assumed that the actions of the Universe were rational, according to a human-rationalized order. In the Greek (Aristotelian) view, the philosophical mind saw truth and perfection in the mind and imposed it onto the Universe. For example, the Greek view on motion might follow this narrative:

- The Gods who made the world are perfect.
- Circles and straight lines are perfect.
- Gods make motion.
- Motion must be perfect because the Gods made it.
- Therefore, motion in the natural world is circles and straight lines.
- By corollary, planets (where Gods live and which must therefore be perfect) move in circles and cannon balls move in straight lines.

The problem with “ways-of-knowing” models of explanation is that all observations are forced to fit the model. The underlying model cannot be changed by evidence. This mode of explanation is resistant to any change in the worldview because new observations cannot alter the underlying models of cause and effect. For example, the idea that motion occurred in straight lines led medieval military engineers to calculate that a cannon ball would rise to a certain height and then fall straight down over a castle wall. However, the cannon balls did not land according to the medieval engineers’ expectations. The Aristotelian “way of knowing” was not able to provide a means to correct the error between what was expected and what happened.

Both of the “received knowledge” and “ways of knowing” modes of explanation satisfy the brain’s goals of completing patterns of cause and effect and of avoiding ambiguity. However, the particular viewpoint of these modes of explanation enhances the brain’s intrinsic tendency to lock onto pre-set biases. Neither the “received knowledge” nor the “ways of knowing” modes of explanation has the capacity to alter the underlying worldview. These modes of explanation are, therefore, limited in their flexibility and capacity to expand their field of knowledge beyond a relatively restricted plane of observation. Both are like looking at the world through a fixed focus lens or, in the most extreme case, closing the lens completely and considering only what is already known and accepted as the extent of relevant knowledge.

Modern skeptical-empirical science. During the Italian Renaissance, Leonardo da Vinci (1452–1519), who was a very good military engineer, tried to solve the problem of the “mortar shells that kept missing.” da Vinci’s approach was radical for his time. He observed that when a mortar was fired, the shell followed a path that was not the one predicted by “perfect” motion. There was no straight-line motion at all! Instead the shell followed the path of a “parabola.” He changed his world’s view based on his experiments and measurements. da Vinci’s mortars began to hit their targets, and the seeds of modern experimental science were planted. The growth of these seeds occurred subsequently in the work of Galileo (1564–1642) and Kepler (1571–1630), whose story we will touch on in the next section.

In the scientific mode of explanation, the fundamental rules are discovered not from assumption or philosophical musing, but rather from careful consideration, measurement, experiment, and analysis of specific, relatively simple cases. Observation is the first step in constructing a model, then testable hypotheses are proposed for relationships within the model. The validity of the model and its hypotheses is tested by making a prediction, performing experiments to test the prediction, making experimental measurements, recognizing that the observer may influence the experiment, accounting for that influence (design of controls) and, then, changing (discarding or revising) the model when the experimental evidence requires a different worldview.

Instead of using a strictly deductive logic that dictated reality from a series of "self-evident" propositions, modern science began by breaking from this tradition and using inductive logic. In the framework of deductive logic, the general proposition exists first. The specific case is logically found starting with the general case and working toward the specific one. In inductive logic, the principles of rational order still stand, but the first step is the consideration of specific cases that are carefully studied, and then the specific case is generalized backward to fundamental principles. In a system of inductive logic, the fundamental rules are usually discovered not from progressing from assumptions, but rather through the questioning of assumptions following measurement, experiment, and analysis.

Modern science uses experimental models and inductive logic to balance the internal formal explanatory models that human imagination formulates. This is an important interaction in modern science. Here we do a brief experiment to explore this balance between a vital human mental capacity and the exploration of the natural world:

2.5 The Gedanken Experiment Is a Thought Experiment

Imagine yourself standing outside a country home on an early spring morning just before sunrise. Take a deep breath and shiver to the taste of the sweet pre-dawn air. Listen carefully to hear the chirping of morning birds. As the sun reaches the horizon, glinting shafts of light reach your eyes. Another deep breath and you feel a peace that comes from a resonance between you and the world at your doorstep. Your eyes close and for a fleeting moment you understand the Universe in its simplest, most basic terms. Savor that moment, for your eyes open again and now you are drawn back to the reality – you are reading the introduction to a book on physical chemistry. If you are mildly perturbed at being returned to this apparently less appealing reality, you have just demonstrated a facility with a key and exquisitely valuable tool in the study of science, the *Gedanken experiment* (thought experiment). The use of thought trips will be of fundamental importance in the approach that this book takes toward understanding biophysical processes. That virtually any student has access to one of the most profound and sophisticated theoretical techniques available to a scientist is an important lesson to learn.

So here you are, just several pages into a textbook on biophysical chemistry and in clear possession of a sense that your imagination, rationally applied, is going to be a necessary tool for our studies. How can this be? Why not just write a text in which the facts about what is known are written down along with an honest appraisal of what has not been discovered yet (so that good fundable grants can be easily identified). In many cases this could be done to some reasonable degree of certainty, but how can you, the reader, be sure which of the facts you will study are (a) highly certain, (b) certain, or (c) just a little certain? Ultimately, all of the information is going to be relayed by a single source, so does that not imply a degree of confidence that you could apply to all the facts in the book. Are not facts facts after all? Is it not true that we know what we know, and the job of science is to simply extend the frontiers of our knowledge forward laying claim to the regions formerly unknown? Questions of this type are fundamental to the study of any scientific endeavor. Although it is unfortunate that the pop-culture treatment of questions about reality and certainty have to some degree reduced these questions to caricatures, they remain at the center of scientific inquiry about the world in which we live.

How do we know what we know about the organization of biological systems at the chemical level? The compelling pictures of macromolecules drawn in subnanometer resolution by a computer are in fact constructions deduced from patterns of scattered x-rays and have never actually been seen by the human eye or even any magnifying instrument that can form an image in the same fashion as the human visual system. How can we be so sure or even have any less doubt that one description is any better than any other? This is not a trivial point and in fact strikes to the heart of the study of knowing that obtained knowledge has any meaning in a real system or world. The study of knowing is called *epistemology* and is the field in which science and philosophy meet. The issues explored in epistemology are fundamental to an understanding of the method of science.

Derived from the Greek roots *episteme* (knowledge) and *logos* (theory), epistemology is essentially concerned with the theory of knowledge (i.e., what is the relationship between structure, origin, and criteria of knowledge). There are many important questions with epistemological roots of tangible and practical as well as intellectual interest to the scientist. For example, fundamental to human knowledge are issues of perception both in terms of sensory perception and misperception, the mode as well as choice of observables and the potential for perceptual illusion. A crucial issue is to define the relationship between the observer and the observed as well as the relationship between the knower (who may be other than the observer) and the object or system known. (The following example of Kepler is appropriate here, since Kepler modeled and proved his laws of planetary motion by using Tycho Brahe’s superb data on the planetary positions. Brahe, although a good observer, had attempted to devise a set of celestial rules that depended on a lunar-geocentric alternative to the Ptolemaic and Copernican theories of the time. His formulation was non-sensical though his observations were accurate and without peer.) The types and kinds of knowledge as well as the degrees of certainty associated with each type of knowledge need exploration. The questions of what comprises truth, whether truth and understanding can be discovered, inferred or calculated, or if truth and understanding are different, are important modern questions of epistemology, and are argued among mathematicians and artificial intelligence, cybernetic, and intelligence-modeling workers. Even the nature, limits, and justification of inferences are important questions of epistemological nature.

2.6 The Beginnings of Modern Science– Kepler and Galileo

Johannes Kepler's approach to and formulation of the laws of planetary motion and Galileo Galilei's exploration of the laws of mechanics and motion mark not only the beginning of modern science but were instrumental in ending Aristotelian ways of knowing. Kepler's planetary laws are fundamentally important not only because of the result but for the shift he made in the premises of his arguments. In *Mysterium Cosmographicum*, Kepler argued that for an argument to be valid, it must be able to pass the test of observable experimentation. He wrote

> What we have so far said served merely to support our thesis by arguments of probability. Now we shall proceed to the astronomical determination of the orbits and geometrical considerations. If these do not confirm the thesis, then all our previous efforts shall have doubtless been in vain. (translation from Koestler, Arthur, *The Sleepwalkers* (New York: Macmillan, 1959, p. 255)

A practicing astrologer, Kepler was born in Germany in 1571. Ironically his interest in knowing the motion of the planets came from a desire for a more precise application of his magical beliefs in the influence of planets and stars on an individual human life. Since he believed in such cause-and-effect phenomena, his view of *causality* was radical for the time. Kepler proposed that the planets moved in their orbits because a force, "spreading in the same manner as light," came from the Sun and kept the planets in motion. Although Newton would show this concept (of anti-inertia) to be wrong, the idea *that things happen because of forces that are mechanical and can be measured* was groundbreaking. This view of causality implied that these mechanical forces could be observed, measured, and used to build a formal geometrical model that would be accurate, in terms of the forces (causes), in predicting the future (effects). It also implied that the application of equivalent forces on an equivalent system would yield equivalent results.

In a paradoxical way, Kepler was one of the pioneering biophysicists. He believed, without any significant evidence, that the mysterious forces of life could be influenced by the positions of the planets and the stars. As an astrologer he accepted that the forces acting between the celestial worlds and the living world of humans had a predictable pattern of correspondence. He believed that he understood the rules that governed the correspondence between human behavior and the positions of the planets. A knowledge of the mechanical rules of astrological forces was assumed. Thus, if he could know the actions (positions and movements) of the celestial bodies, he could apply his knowledge of the rules of interaction (the forces) and predict the effects on life events. By devising an accurate celestial physics he would be able to better understand biological processes. Today we would argue that his astrological rules are not well validated, and in fact we have replaced them with rules of chemistry and molecular cell biology. But the assumption that biological behavior can be understood via an understanding of the physical rules governing the Universe is the fundamental assumption of this book and modern biophysical chemistry.

When Kepler set out to explore the linkage between the physical nature of the Universe and biological systems, he was to attempt to build a formal model of the Universe that could represent the natural system (of planets and stars). What was different from the Aristotelian/Platonic and Ptolemaic tradition was that Kepler's formal model was validated not by what the mind thought should exist, but rather by empirical observation. Thus Kepler used the essential elements of modern scientific model building to develop his celestial mechanics:

(1) He studied a natural system by selecting a set of observables.
(2) The system is described by the specification of these observables and a characterization of the manner in which they are linked.
(3) Although theory may be applied to the problem, contact with the reality of the natural system is made wholly through the observables.
(4) A model can be constructed that formally relates the observables and their linkages such that a good formal model, under the proper circumstances, describes the original natural system to a prescribed degree of accuracy. The behavior of the model and the natural system is thus invariant with replacement of one for the other for a given proper subsystem.

We will explore more fully the nature of model building in Chapter 3. Even though Kepler did not apply his rules of building empirically validated models to his whole system (i.e., the astrologically based biophysical rules of interaction), his approach has been used and refined over the following centuries to replace these astrological rules. Because of his success in establishing a working, useful celestial mechanics, he set in motion the incredibly successful machinery of modern scientific investigation that ironically eventually invalidated his own set of beliefs with regard to the nature of life in our natural world. His failure to successfully advance the end result of a great astrology confirms his seminal contribution to the process of scientific discovery.

At the start of the twenty-first century, it is almost impossible for us to appreciate the icon-shattering nature of Kepler's treatment of astral motion. Kepler proposed and used observables to validate the idea that the path and motion of the planets was neither circular nor constant. The idea that the circle was perfect and that nature would naturally be dominated by a static perfect order was the idealizing belief of the Aristotelian mind that had dominated the intellectual perspectives for almost 2000 years. It is remarkable that Kepler was able to propose that the actual movement of the planets was elliptical and that their speed varied along these paths! It is inconceivable that his physics could have been convincing enough to be accepted if he had not used empirical observation to demonstrate the invariance of his formal model with the natural system. Not only did he establish the usefulness of model making as an effective methodology for discovering knowledge but he laid the groundwork for the dynamical modeling that has characterized modern scientific investigation.

While Kepler was the first biophysicist, Galileo was the first modern physicist. Born in Pisa, Italy in 1564, Galileo, like Kepler, developed and used the modern scientific method or progression of inquiry in his research into accelerated motion and dynamics. Galileo's major contributions were in mechanics and it was in these experiments that he made observations that were mapped to hypotheses expressed as mathematical models and then tested in carefully designed experiments. In his seminal work, *Discourses Concerning Two New Sciences,* he developed the field of mechanical dynamics starting with a definition of acceleration, making careful observations, and then reasoning from these experiences via the application of mathematical tools to arrive at important conclusions regarding motion and the strength of materials and structures. Galileo was the first thorough modern scientist practicing the progression of inquiry as evidenced by his use and emphasis on experimental models and experiment to test physical theory. He clearly recognized that experiment alone, though essential, did not constitute (modern) science since foundational ideas like acceleration transcended (formed the context or background) of laboratory experience. As a historical aside, Galileo is probably best known for his interest and views on astronomy. He gained the attention and displeasure of the inquisition though his firm and very public support of Copernican theory, especially after the appearance a supernova in 1604 which rocked the Aristotelian dogma of the immutable nature of the heavens. Kepler had little influence on Galileo (it was to be Newton who discovered the connections between planetary motion and the law of universal gravitation), but both of these early scientists can be credited with the discovery and application of the modern scientific method or progression of inquiry

2.7 Modern Biophysical Studies Still Follow the Paradigm of Kepler and Galileo

Within a context that will be a constant guide throughout this text, we now explore how the integration of the progression of inquiry into biophysical investigation can be summarized:

2.7.1 Describe the Phenomenon – What Is happening Here? What Are the Emergent Properties of the System?

The knowledge of a phenomenon depends greatly on the way in which it is observed. The choice of observables in the description of a system or in a process operating in a system is one of the most important steps in all science. Interestingly, we will learn that in classical mechanical treatments, the observables themselves are used to describe a system but in quantum mechanical treatments the description of the system depends on certain functions that *operate* on the observables. The use of operator functions to describe the linkages in a quantum mechanical treatment has important consequences and indicates that the observer is very strongly linked to the system under observation hence making independent or non-interactive observation difficult. Observations of systems and the development of methods to predict the behavior of the system under different conditions are important aspects of system science and are used extensively in thermodynamics. Thermodynamic treatments can be very useful when we are describing a system because they do not require a detailed knowledge of what is happening inside the system at all times. Thus certain tools let us use the very useful idea of a *black box*. A black box is a treatment of a system in which only the inputs and outputs are considered and these observables are linked phenomenologically. As a starting point, treating biological systems as black boxes is often a necessary first approximation or abstraction. Mathematical systems theory and thermodynamics are important physical tools at this stage. Studies of cybernetics, chaos, complexity and catastrophe theory, the multiple equilibria that describe ligand–receptor interactions, and the dependence of melting on cooperative interactions are all examples of biological application of these techniques. This is our starting point and we will take up the issue of the study of systems in Chapter 3.

2.7.2 Reduce the Phenomenon to a Systems Description: Identify the Components of a System – Who and What Is Involved? (What Are the Elements?)

Though we could conclude our study with a phenomenological analysis of a biological system, it is important and intellectually satisfying to have a more specific description of the components that make up a system. Many of the laboratory tools used to separate, concentrate, and characterize biomolecules are based on their physical properties such as size, shape, weight, and charge. The techniques sensitive to these properties are based to a great degree on mechanics, kinematics, and the transport phenomena of diffusion and charge movement. For the most part, these ideas are based on classical physics and include centrifugation and sedimentation, electrophoresis, chromatography, and the properties of viscous flow.

2.7.3 Analysis of Structure – What Does it Look Like? What Are the Relationships Between the Components? (What Are the Interaction Rules and What Is the Context of the System?)

In addition to identifying the components we also find it interesting to know how they are arranged and what their structure "looks" like. In biological systems, the description of the structure and properties of the chemical components alone is often not sufficient because these components are arranged into higher order structures such as membranes, organelles, and assemblies such as the electron transport chain. The ordering of these components into structures gives us a double benefit for having learned the physical principles outlined in the previous two sections. This is because the physical processes of transport, equilibrium, and systems analysis underlie the basis of cell biology and the physiology of the organism.

Being visual creatures, we tend to prefer to “see” the objects of our interest as we construct a picture of it in space (this includes linear, planar, three-dimensional, and multi-dimensional space). One of the most important techniques for visualizing structure is through the use of microscopy. Modern microscopy includes not only optical methods that use visible light to examine cells and tissues, but also adds methods that allow imaging of subcellular structures, molecules, and even atoms.

We are able to infer a great deal about the electronic and nuclear structure of molecules because of the interaction between electromagnetic radiation and matter. The theoretical explanation for these interactions, *quantum electrodynamics*, forms the basis of spectroscopy, which has allowed the chemical nature of biological systems to be explored in great detail. An understanding of basic quantum mechanics allows us to explore structure: electronic structure (through ultraviolet and visible spectroscopy); the structure that influences rotational and vibrational movements (through infrared, Raman, and microwave spectroscopy); structural influence on the magnetic spin properties of electrons and nuclei (through electron spin and nuclear magnetic spectroscopy); and the relationship between relativistic motion and structure (through Mössbauer techniques). Finally, because chemical structures can interact with electromagnetic radiation, twisting and scattering it, we gain structural knowledge from the techniques of Rayleigh scattering, polarimetry, circular dichroism, and x-ray crystallography.

We are able to use the physical models expressed in mathematical form along with the measurements of structure obtained from thermodynamics and spectroscopy to explore potential structures with the aid of a computer. A powerful relatively new tool available to the biological scientist is the ability to explore potential energy surfaces as biomolecules interact and find their preferred structure in an abstract state space (the space of possibilities). With the powerful visualization tools now available to most scientists, computational chemistry is a synthesis of art and science that has great theoretical and practical appeal.

2.7.4 *Analysis of Dynamic Function – What Is the Mechanistic or Explanatory Cause of That?*

Most of us are not satisfied to simply look at the system as a taxonomist might, pleased by a properly stuffed and static picture. In general we want to know how and why the system runs. In fact most of our modern interest in the structure of biological molecules and systems is the foreplay related to our passionate interest in function. Thus structure–function studies are the central goal of our investigations.

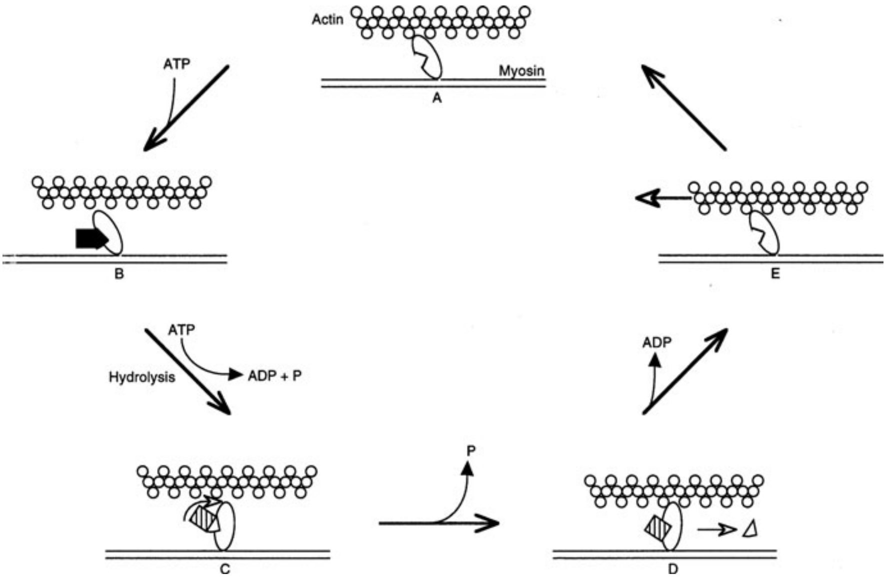


Fig. 2.1 Schematic of the molecular mechanical events in muscle contraction. As described in the text (**a**) the muscular elements are in the "rigor" state. (**b**) The non-bonding interactions between the myosin and actin are disrupted by the binding of ATP to the myosin. (**c**) With hydrolysis of the ATP, the myosin head straightens up. (**d**) When the inorganic phosphate is released, new interactions between the myosin and actin are generated. (**e**) The new intra-chain interactions induce the release of the ADP, which causes the myosin to return to its rigor conformation

The knowledge of how a system works at the molecular level, i.e., muscle contraction (Fig. 2.1), is satisfying because it connects the hidden clockwork to common observation. Mechanistically, a muscle fiber contracts because of the movement of the actin and myosin proteins with respect to one another. The shortening of the muscle follows activation by an electrical switching event, the depolarization of the neuromuscular junction. At the molecular level this shortening occurs as follows: At rest, a myosin molecule's "head" is bound tightly to the actin filament (the *rigor* configuration, from *rigor mortis* – the rigidity of death). The head is released from the actin filament when a molecule of ATP binds to a cleft in the head portion of the myosin thus inducing a conformational change in the myosin and weakening the rigor conformation. Hydrolysis of the bound ATP molecule into P_i and ADP is associated with a sudden straightening of the angle of the head thus causing the head to move nearly 5 nm down the actin fiber. There is a weak bonding interaction between the straight head and a new interaction site on the actin. This interaction causes another small conformational change which causes the P_i to be released from the head. The loss of the P_i releases a "power stroke" in which the myosin head moves back to the rigor angle releasing ADP because of the conformational change and thus pulling the actin fiber 5 nm into a new rigor conformation awaiting a new stroke cycle. Here form and function are elegantly interdigitated.

By exploring the details of molecular interactions, we now have an understanding of how the observables that we first discovered phenomenologically (muscles move) and characterized by a black box treatment with a state function (the force–length relationship of muscle contraction) occur. One of the important aspects of mechanistic studies is that they must give the same result as the black box treatment when adjusted for number, because the observables associated with the black box studies (usually thermodynamics) are generally more accurate. This is because there is a weaker coupling between observer and observed system and hence less opportunity for observer influence over the system (in Chapter 3 more will be said about this important subject).

We usually study systems mechanistically by observing how they change with a perturbation. Because systems at rest are in fact in dynamic equilibrium (this is a concept from statistical thermodynamics, i.e., a system is perturbed or moved away from equilibrium and then relaxes back toward the equilibrium state); the rate at which a system relaxes after being perturbed is the concern of kinetics. Methodologies that measure rates of chemical change provide a window into the response of a system to various perturbations. The molecular understanding of how quickly a system can move depends on a statistical thermodynamic formulation of how likely a perturbed or post-perturbed conformational state will be found. This is because the conformation of molecules and systems are *conservative systems*. Conservative systems are those in which the potential energy and hence the internal energy are related to the mechanical positions of the elements in the system. Since the direction a system moves after being perturbed depends on the energy flow and the energy flow depends on the potential energy gradients, form and function are tied tightly to one another. By studying kinetics we can examine structure from a different perspective and vice versa.

Finally, we can close the loop. Because we are interested in biological systems we are implicitly interested in how the system is integrated or controlled. What a system does and the rate at which it can do something is dependent on the energetics of its form and function. An active system is one that is perturbed away from its equilibrium state and hence its behavior depends on the potential energy surface near the equilibrium state or the next metastable state to which it will move. Controlling the potential energy surface is thus an obvious method of controlling the system. We can therefore speak to substantial quantitative terms about biological control systems if we understand the energetic interactions in a system that influence the potential energy or control surface. Since the cardinal interest of biophysical chemistry is to understand the energy relationships between the molecules and the systems in which they are found, it is obvious that biophysical chemistry is a natural language for understanding biological behavior and the systems that generate that behavior.

Further Reading

A variety of texts approach the interdisciplinary subject of molecular biophysics. Some are physical chemistry texts oriented to biological applications.

Cantor C.R. and Schimmel P.R. (1980) *Biophysical Chemistry, Parts I, II and III*. W.H. Freeman, New York.

Chang R. (2000) *Physical Chemistry for the Chemical and Biological Sciences*, 3rd edition. University Science Books, Sausalito, CA.

Chang R. (2005) *Physical Chemistry for the Biosciences.* University Science Books, Sausalito, CA.

Edsall J.T. and Wyman J. (1958) *Biophysical Chemistry*. Academic, New York.

Eisenberg D. and Crothers D. (1979) *Physical Chemistry with Applications to the Life Sciences.* Benjamin/Cummings, Menlo Park, CA.

Engel T., Drobny G., and Reid P. (2008) *Physical Chemistry for the Life Sciences*. Pearson-Prentice Hall, Upper Saddle River, NJ.

Tinocco I., Sauer K., Wang J.C., and Puglisi I. (2001) *Physical Chemistry (Principles and Applications in the Biological Sciences)*, 4th edition. Prentice-Hall, Englewood Cliffs, NJ.

There are an increasing number of new texts with a biophysics or physical biological approach. These are developing to support the increasing number of students with an interest in quantitative biology and bioengineering.

Beard D.A. and Qian H. (2008) *Chemical Biophysics: Quantitative Analysis of Cellular Systems*. Cambridge University Press, Cambridge.

Daune M. (1999) *Molecular Biophysics: Structures in Motion*. Oxford University Press, New York.

Jackson M.B. (2006) *Molecular and Cellular Biophysics*. Cambridge University Press, Cambridge.

Nelson P. (2008) *Biological Physics: Energy, Information, Life*. W.H. Freeman, New York.

Phillips R., Kondev J., Theriot J., and Orme N. (2008) *Physical Biology of the Cell.* Garland Science, New York.

Sneppen K. and Zocchi G. (2005) *Physics in Molecular Biology*. Cambridge University Press, Cambridge.

Waigh T.A. (2007) *Applied Biophysics: A Molecular Approach for Physical Scientists.* Wiley, Chichester.

The more traditional coverage of physical chemical principles can be found in the following texts. In general these texts are clear and lucid and are a useful place to start an exploration of topics outside the direct biological sphere. The newer texts have an ever increasing amount of biological material.

Alberty R.A., Silbey R.J., and Bawendi M.G. (2004) *Physical Chemistry* 4th edition. Wiley, New York.

Atkins P.W. (2006) *Physical Chemistry*, 8th edition. Oxford University Press, Oxford, New York.

Castellan G.W. (1983) *Physical Chemistry*, 3rd edition. Addison-Wesley, Reading, MA.

Moore W.J. (1978) *Physical Chemistry*, 4th edition. Prentice-Hall, Englewood Cliffs, NJ.

Philosophy and Epistemology

Feynman R.P., Leighton R.B., and Sands M. (1963) *Atoms in Motion, Lecture #1* ***in*** *The Feynman Lectures on Physics*, Volume 1. Addison-Wesley, Reading, MA. (In classic Feynman style he explains the approach to theory, experiment, observable and abstraction.)

Russell B. (1945) *A History of Western Philosophy*, Touchstone/Simon and Schuster, New York. (Fun to read and accessible discourse on philosophy for the non-professional philosopher. Bertrand Russell was a mathematician and his scientific tilt makes the scientist at ease with the subject.)

Muscular Contraction

Amos L.A. (1985) Structure of muscle filaments studied by electron microscopy. *Annu. Rev. Biophys. Biophys. Chem.*, **14**:291–313.

Pollard T.D., Doberstein S.K., and Zot H.G. (1991) Myosin-I, *Annu. Rev. Physiol.*, **53**:653–681.

Rayment I. et al. (1993) Three dimensional structure of myosin subfragment I: A molecular motor, *Science,* **261**:50–58.

Stossel T.P. (1994) The machinery of cell crawling, *Sci. Am.*, **271, 3**:54–63.

Problem Sets

1. A system can be described by listing the (1) overall or "emergent" properties, (2) elements that comprise it, (3) the way the elements are related to one another and to the background or context space, and (4) the characteristics of the contextual space. Write a systems description for several familiar scenarios: (a) a sports event or game, (b) a holiday dinner, and (c) a laboratory experiment.
2. Use a systems analysis to describe the system that Johannes Kepler studied.
3. For each model system developed in this book, make it a habit to write out the systems description whenever you encounter that model. This includes the kinetic theory of gases, thermodynamic systems, the Born model, the Debye–Hückel model, electric circuit models of electrochemical systems, etc.

Chapter 3
Overview of the Biological System Under Study

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_3, © Springer Science+Business Media, LLC 2010

3.1 Hierarchies of Abstraction Are Essential in the Study of Biophysical Chemistry

Biophysical chemistry is a method of simplifying and abstracting from the natural biological system. The foundation theory in biophysical chemistry is the study of the interaction of electromagnetic fields with light and matter. This theory is called *quantum electrodynamics* and unifies the treatment of electrons and electromagnetic fields (see Appendix B). Today, quantum electrodynamics or QED is the best (most accurate) abstraction we have to describe the natural system of chemical interactions in the Universe. It is sufficient to explain all chemical behavior. Therefore the whole of chemistry and biology rests upon its ample foundation stones. Unfortunately as an abstraction, QED is not practical for daily use because it is computationally intractable. We will allude to QED but will not explore it in any great depth in this volume. Instead we will be more concerned with building our chemical biology upon simpler and more accessible abstractions of electronic structure. The most commonly applied abstraction of QED is the *Schrödinger equation*, which treats the electronic structure as if it were a wave. Again, theoretically such a treatment should allow us to understand chemical biology ab initio (from first principles) but is unfortunately too complex and must itself be replaced by a less complete but more accessible abstraction, *molecular orbital theory* (MO). Molecular orbital theory is grounded in quantum theory and as a greatly simplified abstraction is very useful in understanding the structure and properties of molecules. Unfortunately as the size of the molecular system being modeled grows, MO theory suffers severe computational constraints and is most useful qualitatively.

To explore large molecular systems such as proteins and nucleic acids, we again resort to abstraction and assume that their interactions can be modeled classically (i.e., as if they were Newtonian masses connected by springs). Such methods are called *molecular mechanical models*. There is a computational limit to this abstraction as well as we try to account for the movement of hundreds of atoms, even when the movements are considered classically. The solution is to assume that only a small bounded part of our system has *structure.* The remainder of the system is treated as being a *continuum* that can be characterized adequately by a property of the continuum such as the dielectric constant. Depending on the level of our interest in the system, the process of abstraction can be applied even in cases where no detailed structural information is available or is of great importance. These are the levels of *system and subsystem abstraction* in which the physical chemical system can be treated as a *black box* or series of black boxes inter-linked by equations of state. Fundamental biological processes such as organ physiology, evolution, and ecology are often well handled at these levels of abstraction. The hierarchy of abstractions that the biophysical chemist may find useful is listed in Table 3.1. As we will see in the following chapters, we can connect these levels of abstraction using a variety of physical techniques and hence develop a knowledge about the system of interest at differing levels of abstraction as may be appropriate to our needs.

Table 3.1 Hierarchy of abstraction in biophysical chemistry

Quantum electrodynamics
Matrix mechanics or the Schrödinger equation
Molecular orbital theory
Molecular mechanics
Bounded structure/continuum models
Continuum models
Subsystem linked black box models
System black box models

3.2 An Overview of the Cell: The Essential Building Block of Life

The *cell* is the organizational basis for life. The biological process of *evolution* has resulted in a bewildering array of life on the planet Earth. The diversity of successful life forms ranges from solitary cells to highly complex organisms such as humans. All successful living organisms are capable of maintaining a series of interlinked functions whether they are uni- or multi-cellular. These shared properties include

1. An ability to maintain an integrity of self versus otherness – In some cases this is the maintenance of an internal milieu or the development of an immunological defense system.
2. A system for energy production, acquisition, and storage.
3. A system of waste removal and disposal.
4. A capacity to propagate itself and ensure the continuation of its genetic code into the next generation.

We will briefly describe a prototypic eukaryotic cell in order to establish the general principles of cell biology. It should be remembered that many cells such as neurons, platelets, and red cells have developed extremely specialized cellular elements to perform their specific tasks. In spite of this sometimes extreme specialization, all cells share a common ancestral capacity for being pluripotential or for having the ability to become any type of cell.

The prototypic cell is defined in its outer boundary by the cell membrane which encloses the cytoplasm or “cell water” with its various specialized organelles, each in turn surrounded by its own membrane. The genetic code contained in the DNA is found in the nucleus, which is itself surrounded by the nuclear membrane. There is an internal scaffolding inside most cells, the cytoskeleton, which acts both as a support and as a transporting infrastructure.

3.2.1 The Cell Membrane Is a Physical Boundary Between the Cell System and Its Surroundings but This Membrane Is Also Part of the Biological System

The cell membrane that encloses the cell is relatively thick and is on the order of 9–12 nm. These membranes are built from two layers of phospholipid with associated proteins. The lipid membrane has proteins associated with it arranged on the inside and outside faces and piercing through it. Many of these proteins are receptors that provide biochemical interaction with the internal and external environment. Other important protein elements are small molecule channels such as Na^+ and Ca^{2+} channels and transporters such as the glucose and amino acid transporters. Carbohydrates are found on the outer surface of the membrane where they act as recognition sites for cell–cell identification. When cells need to increase the surface area for adsorption of nutrients (such as in the cells of the digestive system), the plasma membrane folds and invaginates to accomplish this goal. Materials can be taken up by the membrane process of *endocytosis* in which the membrane surrounds the object and then is internalized as a vesicle. *Phagocytosis* is a mechanism by which bacteria and viruses are ingested while *pinocytosis* is the ingestion of objects of molecular dimension such as proteins. By *exocytosis* both synthesized substances and materials to be discarded can be expelled from the cell: the vesicular membrane fuses with the plasma membrane and the internal contents of the vesicle are externalized.

3.2.2 The Cytoplasmic Space Is the Matrix of the Intracellular System

The cytoplasmic space is filled with an aqueous sol/gel of soluble proteins and enzymes, the *cytosol*. There is no distinct structure to the *cytosol* when it is examined under the light or electron microscope. The cytosol is the site of a wide array of metabolic pathways including glycolysis, the pentose phosphate shunt, fatty acid, and glycogen synthesis, and the binding of tRNA to amino acids. Within the cytosol are found energy and building materials, fat droplets, glycogen granules, and protein bodies. The cytosol is traversed by a filamentous network of *microtubules* and *microfilaments* which form the *cytoskeleton*. These networks serve to support and provide mechanical rigidity to the cell. In addition movement of the cell and its organelles is mediated by the cytoskeletal system. The major components of the cytoskeleton are listed in Table 3.2.

The cytosol is the location of protein manufacture (Fig. 3.1). The site of protein synthesis is the *ribosome*, a structure formed from ribosomal RNA and protein (in eukaryotes the ratio is 40:60%, respectively; this ratio is reversed in bacteria). It is at the ribosome that the nuclear code which was transcribed from DNA into

Table 3.2 Components of the microtubules and microfilaments

Name (component)	Molecular weight (kDa)	Size (diameter) (nm)	Organization
Microfilaments (cellular shape and movement)		8	Two stranded helix
Actin	42		Globular shape
Intermediate filaments (mechanical stability)		10	Coiled α-helix coils
Keratins	40–70		Fibrous rod: α-helix
Vimentin	54		Fibrous rod: α-helix
Desmin	53		Fibrous rod: α-helix
Neurofilaments	60–130		Fibrous rod: α-helix
Nuclear lamins	65–75		Fibrous rod: α-helix
Microtubules (cilia, flagella, anchor membrane-bound organelles)		30	Helical array of 13 tubulin units
Tubulin (α and β)	50		Globular shape

mRNA is read and translation from genetic code into a primary sequence peptide occurs. Prokaryotic and eukaryotic ribosomes have two subunits, large and small which depend on the presence of Mg^{2+} to remain connected. (Ribosomes are characterized by Svedberg units, *S*, which is the sedimentation coefficient. In bacteria the ribosome is 70S and the subunits are 50S and 30S while in eukaryotes, the ribosome is 80S with subunits of 60S and 40S.) The small subunit contains sites for the binding of the amino acid-charged tRNA molecule, the tRNA that was associated with the amino acid just incorporated into the new peptide, and the mRNA strand which is the recipe for the primary structure of the protein. The large subunit contains the machinery to catalyze the formation of the peptide bond. The ribosome essentially moves down the mRNA strand like a ticker tape reader, assembling the appropriate amino acids in the correct order and reeling off a new peptidyl chain. The protein is released when a stop codon in the mRNA is reached. Most ribosomes in the cell are aggregated into polysomes though occasional single ribosomes or monosomes can be found. The free ribosomes release proteins into the cytosol. Frequently the ribosomes are attached to an internal membrane system, the *endoplasmic reticulum* (ER), which is called "rough ER" when ribosomes are attached and "smooth ER" when there are no ribosomes. Ribosomes that are attached to the ER feed the nascent protein into the membrane channel formed by the ER. The proteins that enter the ER are mainly produced either for export or secretion or as membrane proteins. Most proteins destined for membranes have a special peptide sequence, the *signal sequence* which directs the protein into the membrane. The signal sequence is usually comprised of hydrophobic amino acids thus making insertion into the lipid membrane energetically favored.

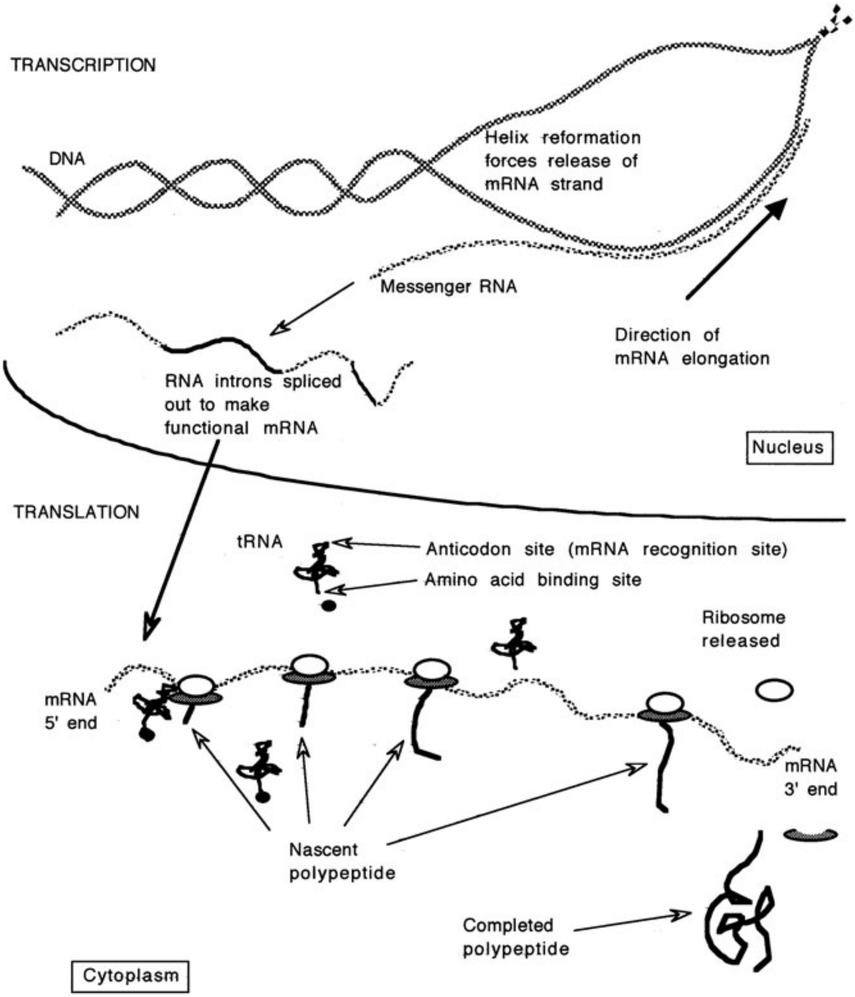


Fig. 3.1 Cartoon of protein production from the gene to the ribosome

3.2.3 The Organelles Are Subsystems that Are Found Within the Cytoplasmic Space but Have Unique Environments and Are Therefore Complex Physical Systems

Organelles are membrane-bound compartments designed for specific purposes. Generally the organelles contain specialized physical environments.

Endoplasmic reticulum. The endoplasmic reticulum has been described in relation to protein synthesis. The ER is an extensive organelle comprised of narrow cisternae, vesicular apparatus, and tubules. The inner diameter of the ER is

20–30 nm. It has a bilayer membrane structure that is 7–10 nm thick. The ER is the site of synthesis of sterols, phospholipids, triglycerides and hence assembly of the membrane. It is the site of detoxification in the cell (via enzyme systems like the P-450 system) and is the site of blood glucose production via the enzyme glucose-6-phosphatase (Table 3.3). In certain specialized cells the ER becomes specialized and controls intracellular transport and supply of ions. An excellent example of this is the sarcoplasmic reticulum in myocytes and cardiocytes where Ca^{2+} is sequestered and controlled in order to govern myofibril contraction.

Table 3.3 Enzymes of the endoplasmic reticulum

	Space in which enzyme is active	
Enzyme	Cytosol	Lumen
Cytochrome b_5	√	
Cytochrome b_5 reductase	√	
NADPH-ferrihemoprotein reductase	√	
Cytochrome P-450	√	√
Cytochrome P-450 reductase		√
ATPase	√	
5’-Nucleotidase	√	
Nucleoside diphosphatase		√
GDP-mannosyl transferase	√	
Glucose-6-phosphatase		√
β-Glucuronidase		√
Hydroxymethylglutaryl-CoA reductase	√	
Enzymes of steroid synthesis	√	
Enzymes of cholic acid synthesis	√	

Golgi Apparatus. The *Golgi apparatus* or *dictyosome* is an organelle of multiple stacked cisternae of the same membrane type as the ER. This is not because they are the same structure but rather because there is a very rapid exchange via a vesicular transport system between the ER and the Golgi so the membrane components of the Golgi are essentially in equilibrium with those of the ER. The Golgi complex is polar with a *cis* and *trans* face and an intermediate *medial* compartment. The *cis* face, which faces the nucleus, receives vesicles from the ER containing newly synthesized proteins and lipids. Proteins arriving at the *cis* cisternae are sorted, new proteins make their way through the Golgi where they undergo glycosylation and final sorting in the *trans* compartment (which faces the plasma membrane). The processed protein is packaged in a variety of vesicles for delivery throughout the cell including secretion from the cell. Proteins that belong to the ER are repackaged at the *cis* cisternae and are returned to the ER. The membrane of the *trans* Golgi is intermediate in dimension between the ER and the plasmalemma.

The Lysosomes. Lysosomes are vesicles whose principal function is intracellular digestion and degradation of biochemicals. Lysosomes are acidic with a pH of 5 which is generated by a ATP-driven proton pump. This acidic environment is necessary for the optimal activity of the hydrolytic enzymes that are contained in these unique organelles. The lysosomes contain lipases, nucleases, proteinases, glycosidases, phosphatases, and sulfatases. Primary lysosomes are formed directly from the Golgi complex and are about 50 nm in diameter. They are clathrin coated upon formation. Following removal of the clathrin, they can fuse with other membranous structures to form larger secondary lysosomes of 300–500 nm diameter. The secondary lysosomes are the site of hydrolytic digestion. The membranes of the lysosomes are of typical 7–10 nm thickness. The comparative structure in plant and fungal cells is called a *vacuole*.

Microbodies. Microbodies are specialized vesicles in which oxidation of amino acids, uric acid, phenols, alcohols, and fatty acids takes place mediated by flavin oxidases. In animals these vesicles are called *peroxisomes* and in plants they are *glyoxysomes*. Their membrane is 5–8 nm thick and the vesicles are 200–1500 nm in diameter. All microbodies contain catalase to degrade H_2O_2 which is a by-product common to all of the flavin oxidases. Thus microbodies have a uniquely harsh oxidative environment comparable to the harsh pH environment of the lysosomes.

The Mitochondrion. According to the endosymbiotic hypothesis this organelle is the descendent of ingested aerobic bacteria. The mitochondrion is the site of the electron transport chain, the citric acid cycle, and fatty acid catabolism (Table 3.4). The ATP necessary for cellular function is generated in the mitochondrion. Mitochondria are oblong, about 0.5 × 1.5 μm. They have a dual membrane system and there are two compartments in the organelle. The outer membrane separates the mitochondrion from the cytoplasm and is similar to the membrane of the ER in terms of thickness, density, and lipid:protein ratio. The inner membrane is similar to the cytoplasmic membranes of bacteria. Inside the inner membrane

Table 3.4 Selected enzymes of the mitochondria

Matrix	
2-Oxoglutarate dehydrogenase complex	3-Oxoacid-coA transferase
4-Enoyl-coA reductase	Acid-coA ligase
Aconitate hydratase	Alcohol dehydrogenase
Aldehyde dehydrogenase	Citrate synthase
Fumarate hydratase	Malate dehydrogenase
Nucleoside-triphosphatate adenylate kinase	Phosphoenolpyruvate carboxylase
Pyruvate dehydrogenase complex	Pyruvate dehydrogenase kinase
Pyruvate dehydrogenase phosphatase	Superoxide dismutase
Inner membrane	
3-Hydroxybutyrate dehydrogenase	3-Hydroxyacyl-coA dehydrogenase
Choline dehydrogenase	Succinate dehydrogenase
Acyl-coA dehydrogenase	NADH transhydrogenase
H^+-transporting ATP synthase	Cytochrome *c* oxidase
Outer membrane	
Amine oxidase	Cytochrome b_3 reductase
NADH dehydrogenase	
Intramembrane space	
Glycerol-phosphate dehydrogenase	Adenylate kinase
Nucleoside diphosphokinase	

is the matrix. The inner membrane is the site of the electron transport chain as well the enzymes of fatty acid β-oxidation. A wide variety of transport systems are also found in association with this membrane. An *elementary particle* called the F_1-*ATPase* is found on the matrix side of the inner membrane and this is the site of ATP formation driven by the H^+ gradient that is generated by the electron transport chain. If these elementary particles are detached or absent as in the brown fat of hibernating bears, the mitochondrion makes no ATP but generates heat instead. Within the matrix are found a wide variety of soluble enzymes as well as mitochondrial DNA and 70S ribosomes. Though the mitochondrion contains its own apparatus for protein production, the majority of mitochondrial proteins are made in the nucleo-cytoplasmic compartment of the cell and transported to the mitochondria.

3.2.4 The Nuclear Space Is an Intracellular Space that Is Separated from the Cytoplasmic Space Because of the Systems Interactions

We have defined the eukaryotic cell as a cell with a separate nucleus in which is found the DNA that directs the cell. The nucleus is not a safebox in which the genetic material is stored until cell replication calls for the duplication of the DNA. The DNA oversees the production, differentiation, and growth of the cell in real time. The nuclear membrane must be specially adapted to interact via RNA with proteins and protein products and to monitor the overall state of the cell responding to both endogeneous and exogeneous inputs. The nuclear membrane is a double membrane structure with two leaflets each 7–8 nm in diameter that enclose cisternae of 10–30 nm in dimension. The inner and outer membranes of the nuclear envelope are defined both topographically and also by the specific proteins that they contain. The leaflets of the nuclear membrane are derived from the endoplasmic reticulum during mitosis. The outer leaflet of the nuclear membrane is contiguous with the ER and shares certain protein and enzymatic activities with rough ER. Ribosomes on the outer leaflet make proteins that are transported into the space between the outer and inner leaflets, the perinuclear space. The perinuclear space is contiguous with the inner lamina of the ER. The inner leaflet contains proteins that anchor a meshwork of proteins called *nuclear lamins*. These interconnected proteins, consisting of intermediate filament proteins, polymerize into a two-dimensional latticework that shapes and supports the nuclear envelope. This structure is called the *nuclear lamina*. Connections are made between the “nucleosol” and the cytosol and its organelles through *nuclear pores*. These openings are 50–70 nm in diameter and the membrane leaflets are fused at the edge of the pores. The openings are lined with an octagonal system comprised of 10–18 nm subunits, the *nuclear pore complex*. These structures are highly complex with a molecular mass of 125,000,000 Da and a composition of approximately 100 proteins. These channels have differential effect on molecules of different sizes with apparently free permeability to molecules of 5000 Da or less, impermeability to molecules over 60,000 Da and an permeability inversely proportional to molecular weight for molecules between these limits. The dimension of the nuclear pore complexes can be estimated by assuming them to be a water-filled pore and measuring the diffusion rate through them. Such treatment indicates a cylindrical channel of 9 nm diameter and 15 nm length. This is an example of how the physical concept of diffusion and chemical transport that we will subsequently study can be used to hypothesize a physical structure in a biological system. Clearly other mechanisms must be available for the transport of the large mRNA and proteins into and out of the nucleus, and the nuclear pore complexes also contain sites for active transport of macromolecules through the nuclear membrane.

The nuclear DNA is contained within the nuclear envelope in a mixture of protein, DNA, RNA, and lipids. The dry weight of the nucleus is approximately 80% protein, 12% DNA, 5% RNA, and 3% lipid. The majority of the nuclear proteins are associated with the DNA and are highly positively charged peptides called *histones*. Other non-histone chromosomal proteins are also found associated with the DNA and play important roles in the regulation of DNA activity. The universally known double helix nucleic acid secondary structure is explored in Chapter 18 but DNA has an important supramolecular organization. Histones are positively charged because of their high content of lysine and arginine and associate strongly with the highly negatively charged DNA double helix. There are five histones: H1, H2A, H2B, H3, and H4. The last four of these are called the *nucleosomal histones.* These are small peptides containing 102–135 amino acids. The H1 histone is a 220 amino acid protein. The nucleosomal histones associate into an octamer with stoichiometry of $H2A_2H2B_2H3_2H4_2$ which is comprised of two (H2A–H2B) dimers associated with a tetramer of $H3_2H4_2$. The nucleosomal histones thus form core particles around which the DNA double helices wind to form *nucleosomes*. One hundred and forty six bases pairs wind around an octamer 1.75 times to form the nucleosome which is then connected by a linking region of approximately 60 base pairs to another nucleosome. This chain of nucleoprotein measures about 10 nm in diameter and has the appearance of beads on a string. The winding of the DNA double helix around the histone octamer is constrained because the radial wrapping of two colinear strands will cause the inside curve to be compressed and the outside curve to be expanded. (Imagine a spring wrapped around a ball.) *A*denine–*T*hymidine pairs are more easily compressed than *G*uanine–*C*ytosine pairs, and so when AT pairs are found on the inner minor groove, there is preferential connection of the histone octamer to these locations. The chromatin is very infrequently seen as a 10 nm strand of beads on a string. Generally the nucleosomes wrap around the H1 histone protein and assemble into a 30 nm fiber. The non-histone chromosomal proteins play an important role in the controlled unwinding of the DNA which is necessary for it to be transcribed or replicated.

Finally, in addition to the highly organized membrane, pores, and chromatin, the nucleus contains the nucleolus, which is the site of ribosomal RNA production. This structure which appears in metabolically active cells is not membrane bounded but is organized into a fibrillar center comprised of non-active DNA, a fibrillar component which is rRNA under transcription and a granular component which is composed of mature ribosomal precursor particles. The size of the nucleolus varies depending on the cell cycle and the activity of the cell but can become large enough to occupy 25% of the volume of the nucleus.

3.3 Control Mechanisms Are Essential Process Elements of the Biological State Space

Biological systems are very complicated with a large variety of processes that ensure the survival and perpetuation of the living being. These go on in a coordinated fashion both in parallel and in proper sequential order. The coordination of these processes requires a series of control mechanisms that are the province of systems science and the field of cybernetics. Much of the study of physiology is aimed at discovering the specifics for applying these basic control mechanisms to particular biological systems. In general, most control mechanisms can be considered either in thermodynamic (state functions) or in kinetic terms. An example of control in a compartmentalized system is the need to ensure proper targeting of proteins made in a general pool such as the ER into the correct compartment such as a lysosome, peroxisome, plasma membrane, or mitochondrion. The use of signal sequences on proteins is a method of targeting. As Table 3.5 shows, the interactional energies

Table 3.5 Selected signal sequences

Function	Peptide	Forces
Go to ER	$^{+}$H$_3$N-Met-Met-SER-Phe-Val-SER-Leu-Leu-Leu-Val-Gly-Ile-Leu-Phe-Trp-Ala-THR-*Glu*-Ala-*Glu*-Gln-Leu-THR-**Lys**-Cys-*Glu*-Val-Phe-Gln	*Hydrophobic* *Anionic*/**cationic** groups HYDROGEN BONDING
Go to mitochondria	$^{+}$H$_3$N-Met-Leu-SER-Leu-**Arg**-Gln-SER-Ile-**Arg**-Phe-Phe-**Lys**-Pro-Ala-THR-**Arg**-THR-Leu-Cys-SER-SER-**Arg**-Tyr-Leu-Leu	Sequence forms amphipathic secondary structure with polar residues on one face and the hydrophobic residues on the other face
or		
	$^{+}$H$_3$N-Met-Leu-**Arg**-THR-SER-SER-Leu-Phe-THR-**Arg**-**Arg**-Val-Gln-Pro-SER-Leu-Phe-**Arg**-Asn-Ile-Leu-**Arg**-Leu-Gln-SER-THR	
Go to nucleus	Pro-Pro-**Lys-Lys-Lys-Lys-Arg-Lys**-Val	**Cationic** groups
Go to peroxisome	SER-**Lys**-Leu	Sequence of H-bonding, cationic, hydrophobic residues
or		
	SER-**Lys**-Phe	

Underline = aliphatic (hydrophobic) amino acids; Italic = anionic amino acids; Bold = cationic amino acids; Small caps = hydrogen bonding amino acids.

between the signal peptide and the organelle depend on the charge, hydrophobicity, and probably the secondary/tertiary structure. Thus forces that determine the control events, or the potential energy control surfaces in the cell, are able to be paradigmatically understood in biophysical terms. Failure of correct delivery due to errors in the targeting systems leads to a wide variety of biochemical diseases including the storage diseases such as Tay-Sachs, Gaucher's disease, and the mucopolysaccharidoses. Some of these diseases are summarized in Table 3.6.

Table 3.6 Examples of diseases caused by incorrect intracellular traffic

Disease	Clinical picture	Trafficing error	Biochemical result
Mucolipidoses II (I-cell disease)	Progressive mental retardation Stunted growth, liver, heart, and spleen disease. Death by age 10	Abnormal or missing GlcNAc-phosphotransferase in Golgi fails to produce mannose-6-phosphate labeling enzymes to be imported to lysosomes. Instead the enzymes are excreted out of the cell	Without the proper complement of catalytic enzymes, the lysosome accumulates undigested proteins and swells. This leads to the characteristic inclusion bodies (I-cells)
Zellweger's	Severe degeneration of the brain, liver, and kidneys. Death in 6 months of birth	Mutation in a peroxisomal assembly protein prevents normal importing of peroxisomal enzymes	Peroxisomes are "empty" with absent cellular metabolism of very large chain fatty acids, bile acids plasmalogens (ether-glycolipids needed in membranes and myelin formation

3.4 Biological Energy Transduction Is an Essential Process that Provides Energy to Ensure the High Degree of Organization Necessary for Life

What will become quantitatively clear in Chapter 11 when we study entropy is that the essential order of living things runs against the natural redistribution-to-the-mean of entropy. Living beings must utilize enormous amounts of energy in order to become and remain organized and therefore alive. In following sections we will begin our quantitative treatment of energetics but here present the big canvas on which our further studies will provide detail.

Thermodynamics is the accounting process for keeping track of the ebbs and flows of energy, matter, and order in a system. Life depends on energy and life is characterized by change and order. We use thermodynamic measures extensively in the study of biological systems for these reasons. Biological energy resource management includes defining energy: sources, acquisition, transformation, storage, and distribution. Because energy and biology, and energy and thermodynamics are so closely related, we will use the example of biological energy management to herald our discussions of thermodynamics.

The solar energy from the Sun is the primary source for most of the energy used in biological systems. The common pathway for energy utilization in all cells studied is by way of the chemical compound *ATP* or *adenosine triphosphate*. Biological energy thus depends on photochemistry. The materials available to build this photochemical system constrain the design to a substantial degree. First of all, the Earth's Sun is a G2 dwarf star with a core temperature of approximately 15 million K and a surface spectral maximum intensity of 5000 Å which accounts for its yellow color. The spectral distribution of the Sun can be treated as a blackbody radiator as shown in Fig. 3.2a. The interaction of electromagnetic energy (photons) with the electronic structure of atoms is the essential physics in chemistry. It is this ability of electrons to interact with light that allows solar–biological energy transfer to exist. The essential step in the whole process is the ability of an electron to absorb (or emit) a photon of certain energies thus enabling the electron to use this energy to move to a variety of energy states. The molecular structure of a chromophore such as chlorophyll has evolved to create an electronic bound state that can interact with a particular energy photon (Fig. 3.2b). This is a glib but essentially accurate summary abstraction of photochemistry. In certain energy states interactions with electrons in other atoms give rise to chemical interactions. In some cases the electron can undergo actual transfer from one molecule to another. Since the electron is charged and possesses a certain energy, such a *charge transfer* can be viewed as associated with either storage or release of energy. The storage or release of energy associated with charge movement is the essential element in electrical work, and the apt analogy of a battery or a fuel cell comes to mind. We can now see the linkages of *electrochemistry* in our system. The system of biological energy transduction is best viewed as a photochemical–electrochemical system.

Since the primary chemical elements in biological systems are carbon, hydrogen, nitrogen, and oxygen, we would correctly anticipate that these molecules play a major role in energy transformation. The transfer of electrons to one chemical species requires the donation of those electrons from a second species. The species that receives the electrons is said to be *reduced* and the species losing the electrons is *oxidized*. All charge transfer reactions occur in pairs in this fashion and are called reduction–oxidation couples or *redox* pairs.

This is nomenclature that many students find confusing. The antidote to this confusion is to remember that all redox reactions occur in pairs and there is a syntax to these descriptors. Understand the grammar and you will not be so easily confused. A *reducing agent* donates electrons to the chemical species being reduced. The reducing agent is thus oxidized. An *oxidizing agent* accepts electrons from the chemical species being oxidized. The oxidizing agent is thus reduced.

Different molecules have differing tendencies to accept or donate electrons. This tendency is related to the final potential energy of the pair and is measured by the

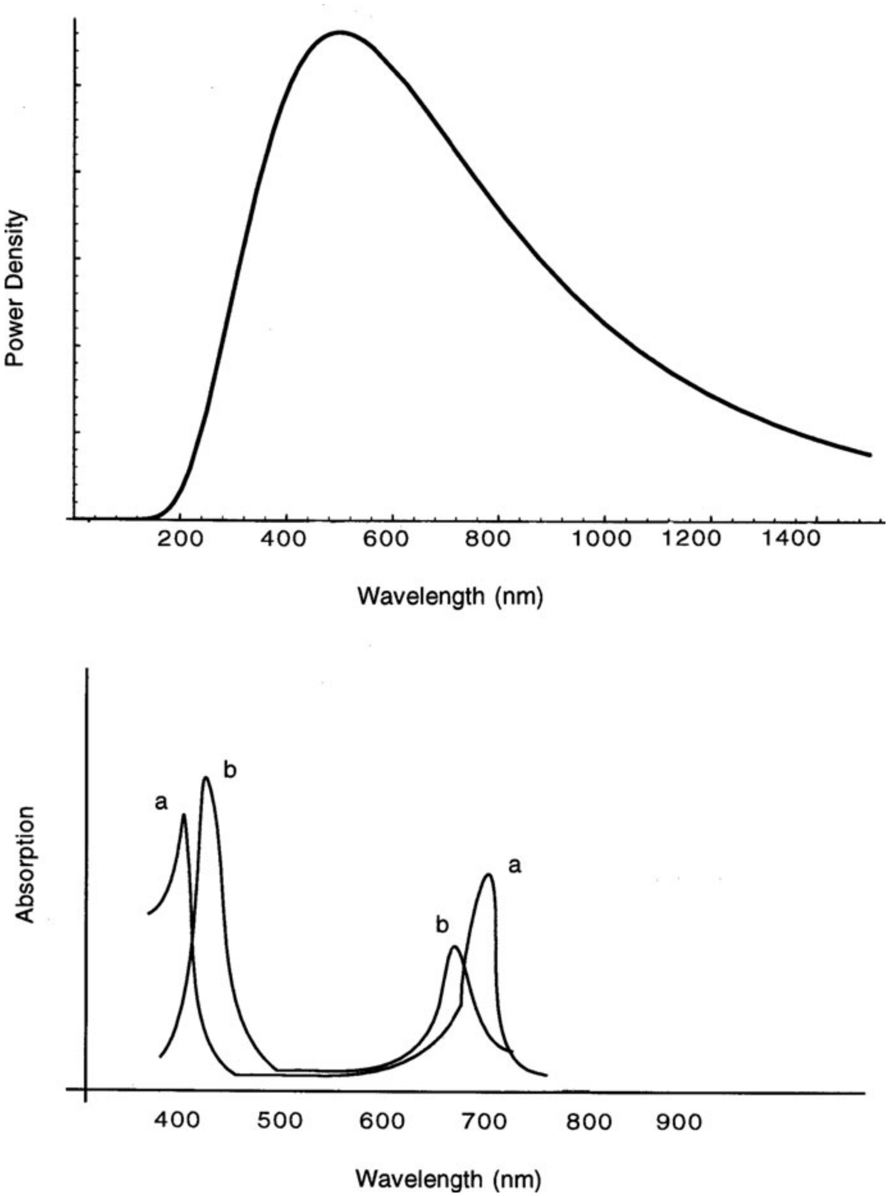


Fig. 3.2 (**a**) top: Spectral distribution of the Earth's Sun is that expected from a blackbody radiator at 5800 K. (**b**) bottom: Chlorophylls a and b have peak light absorbance in the region of maximum spectral power density for a 5800 K blackbody. The chlorophylls are green because they absorb most of the yellow and blue light from the Sun leaving green frequencies to be reflected to the observer's eye

thermodynamic property of the reduction–oxidation potential. The rate of electron transfer is a matter of kinetics and also is important to the likelihood of observing the charge transfer event. In biological systems the strongest oxidizer is usually oxygen. Oxygen therefore has a strong tendency to accept electrons from other molecules often combining with the reducing species to form an oxide. Electrons are negatively charged and in many cases may be transferred in an electroneutral fashion with an accompanying positive charge. In aqueous systems substantial numbers of positively charged H^+ are readily available, and a single electron transfer is often accompanied by the associated acceptance of a proton by the reduced molecule (which acts like a Bronsted base). In water, a two-electron transfer is often accomplished by a hydride transfer ($H^\bullet$). Because so many bio-organic reactions take place in aqueous solution, it is commonly appreciated that reduction occurs when a molecule gains hydrogen.

The overall reactions in which biological charge transfer results in energy transduction use oxygen and carbon as the primary coupled redox pair. The reduction series for carbon and for oxygen are listed in Fig. 3.3. In plants, the fully oxidized form of carbon, CO_2, is reduced primarily to the polyalcoholic sugars that serve to store the photic energy from solar radiation in the chemical bonds of the sugar molecules. The biochemical processes of carbon reduction (or fixation as it is sometimes called) utilize the electron transfer molecule, NADPH. The source of the

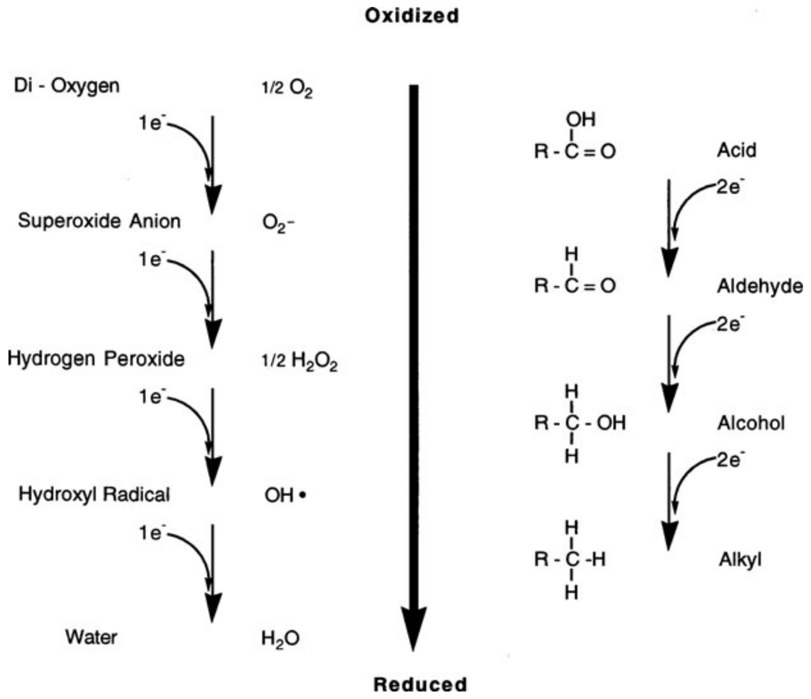


Fig. 3.3 The electrochemical reduction series for carbon and oxygen

electrons that will reduce the CO_2 carbon to an alcohol or hydrocarbon redox state is either H_2O or in some specialized cases H_2S. There are energetic constraints on the use of electrons from water because of the intense electronegativity of oxygen compared to sulfur and these effects are reflected in the reduction potentials of the molecules as can be seen in Table 3.7. With the addition of energy, H_2O can be used as an electron source for carbon fixation, but this requires that a photochemical apparatus be constructed capable of converting light to chemical energy. The removal of four electrons from two water molecules (which is the fully reduced form of oxygen) leaves the fully oxidized form of oxygen, diatomic oxygen, as a by-product. An electrochemical system capable of controlling the energetic electronic species is then necessary to guide formation of the correct compounds and avoid inadvertent energy waste. Furthermore, the oxidized water is highly reactive and steps must be taken to prevent adventitious toxic effects on the organism from the reactive oxygen species produced by the photosystem. The system that performs this light-driven carbon reduction with water oxidation is the photosynthetic reactions found in the thylakoid organelles of plants.

Table 3.7 Standard reduction potentials for biochemical reactions

Reaction	n	E_0 (V) at pH 1	E'_0(V) at pH 7
Succinate + CO_2 → α-ketoglutarate	2		–0.67
Acetate → Acetaldehyde	2	–0.581	
Ferredoxin (Fe^{3+}) → (Fe^{2+})	1	–0.432	
Fe^{3+} → Fe^0	3	–0.036	
$2H^+$ → H_2	2	0.00	–0.421
NAD^+ → NADH	2	–0.320	
$NADP^+$ → NADPH	2	–0.105	–0.324
Horseradish peroxidase (Fe^{3+}) → (Fe^{2+})	1	–0.271	
Glutathione (ox) → Glutathione (red)	2	–0.23	
Acetaldehyde → ethanol	2	–0.197	
Pyruvate → Lactate	2	–0.19	
Fumarate → Succinate	2	0.031	
Myoglobin (Fe^{3+}) → (Fe^{2+})	1	0.046	
Cytochrome *b* (Fe^{3+}) → (Fe^{2+})	1	0.07	
Dehydroascorbate → Ascorbate	2	0.08	
Ubiquinone (ox) → Ubiquinone (red)	2	0.10	
Cytochrome *c* (Fe^{3+}) → (Fe^{2+})	1	0.254	
Cytochrome a_3 (Fe^{3+}) → (Fe^{2+})	1	0.385	
Fe^{3+} → Fe^{2+}	1	0.771	
$\frac{1}{2}O_2 + 2H^+ \rightarrow H_2O$	2	1.229	0.816

Alternatively, other organisms can feed on the reduced carbon molecules produced by the plants and retrieve the energy stored in the carbon bonds by the reverse process. By withdrawing electrons from the hydrocarbon or alcoholic species the carbon atom is moved toward its oxidized form, CO_2. In so doing, non-photosynthetic energy systems shuttle the electrons into a cellular system called the *electron transport chain*. The electron transport chain generates a common energy specie, ATP, which is used by the organism to perform the wide variety of work needed to order and organize the living system. The electron transport chain essentially guides electrons down an energy gradient to O_2 finally yielding free energy and fully reduced oxygen (H_2O).

Why is energy production and transfer so important? We know from experience that objects never spontaneously become hotter and that everything tends to grind to a halt unless maintained in a working state. Everyone with a desk knows that the natural tendency of a desktop is toward disorder unless we persistently straighten it up. Few systems are more organized than even the simplest biological systems. How can this order and hence life be sustained? The answer will be quantitatively evident following our discussion of free energy; but experience tells us that with the input of enough energy and appropriate work we can maintain most systems in a properly ordered fashion. So the organization on which life depends is itself dependent on the availability of energy and its proper application, work. We will see over and over again that by following the energy flow of a system we will gain tremendous practical insight into its inner workings.

The net system behavior in photosynthesis can be expressed in deceptively simple terms:

$$6CO_2 + 6H_2O + \text{sunlight} \longrightarrow C_6H_{12}O_6 + 6O_2$$

This statement is wholly correct; but a little common sense reflection will remind even the most optimistic among us that a flask of CO_2 and H_2O sitting in the Sun will not spontaneously generate sugar and oxygen even if enough energy in the form of light is pumped into the system.

Similarly, the net system behavior in non-photosynthetic metabolism can be expressed in quite simple terms:

$$C_6H_{12}O_6 + 6O_2 \longrightarrow 6CO_2 + 6H_2O + \text{energy}$$

This statement is a little more consistent with our experience because we all have seen a wood fire burn in a hearth and we appreciate that a heat engine powered by the oxidation of carbon fragments (wood, gasoline, or alcohol) can do work. It is not such a leap then to imagine a biological machine capable of burning a carbohydrate and doing work of a biological nature. Perhaps we can imagine a biological machine capable of running in reverse to make the carbon fuel source powered by the Sun. This solar machine would then be a candidate for our photosynthetic machine. The overall black box approach is shown in Fig. 3.4.

Now that we are ready to begin considering what machinery might be a candidate to perform our overall systemic needs, it is good systems analysis practice to know what materials are available. Determining the elements is the next logical step. Our knowledge of model making tells us that our job of identification will be much simpler if we can find certain recurrent themes into which the players will group as “families” with certain common features. Looking ahead we will also anticipate that molecules sharing common identifying features may well also share common

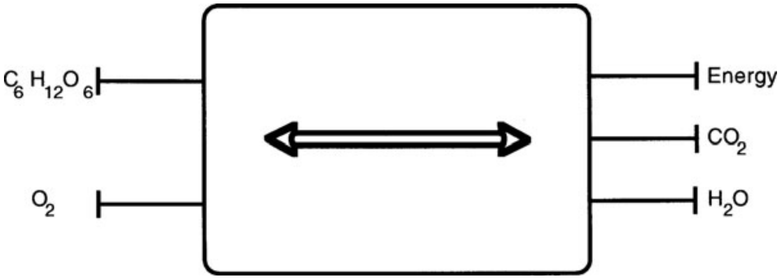


Fig. 3.4 Black box model of a hydrocarbon engine and a solar-powered energy storage device

relationships (e.g., functions and mechanisms) of action. We will focus on photosynthesis, the process of making a reduced carbon molecule from CO_2 using electrons from water as the reducing agent. This process gives us an energy path into which we will insert various chemical species as they are identified. Follow the sketched outline of photosynthesis in Fig. 3.5 in which the energetics and associated components are summarized. Electrochemically, H_2O is strongly oxidizing and therefore holds tightly to electrons. It has a reduction potential of +820 mV. In order to separate electrons from water and release atomic oxygen, energy is required. This excitation is provided in photosystems I and II by the absorption of photons by the chromophores chlorophyll *a* and *b*. The energy trapped by the chlorophyll is provided to a reaction center (P680) in the 600 kDa decapeptide photosystem II. Excited electrons at the P680* center are transferred first to pheophytin and then to a quinone, plastoquinone. In order to regenerate the reduced P680 center, electrons are stripped from water with the concomitant release of oxygen and the production of a proton gradient. Manganese is a necessary cofactor in this step. By bringing the proton gradient into the discussion we are jumping ahead and anticipating the question of structure/function. This is because gradients of charged species such as protons require separation of charges and this separation requires a surface, usually a membrane in biological systems. We have begun the process of determining the structure of our system. In summary, photosystem II drives a reaction that is thermodynamically uphill by using energy extracted from photons of sunlight. The energy is stored as charge and converted into chemical energy by utilizing the energy differences of separate phases, our topic of Chapter 13.

The photonic energy trapped in photosystem II is made available by generating reduced plastoquinone (QH_2) whose electrons are then funneled through another protein complex, the cytochrome *bf* complex. Cytochrome *bf* complex couples photosystem II to photosystem I. Electrons are shuttled from photosystem II through an 11 kDa protein, plastocyanin, and then into photosystem I. As the reducing power stored in QH_2 flows through the cytochrome *bf* into photosystem I, a proton pump generates a proton gradient of $2H^+$ per oxidized QH_2. These proton gradients provide the energy source used to generate ATP via an enzyme complex called ATP synthetase.

Photosystem I is another transmembrane polypeptide complex (>800 kDa) made from 13 polypeptide chains and having much in common with photosystem II. In

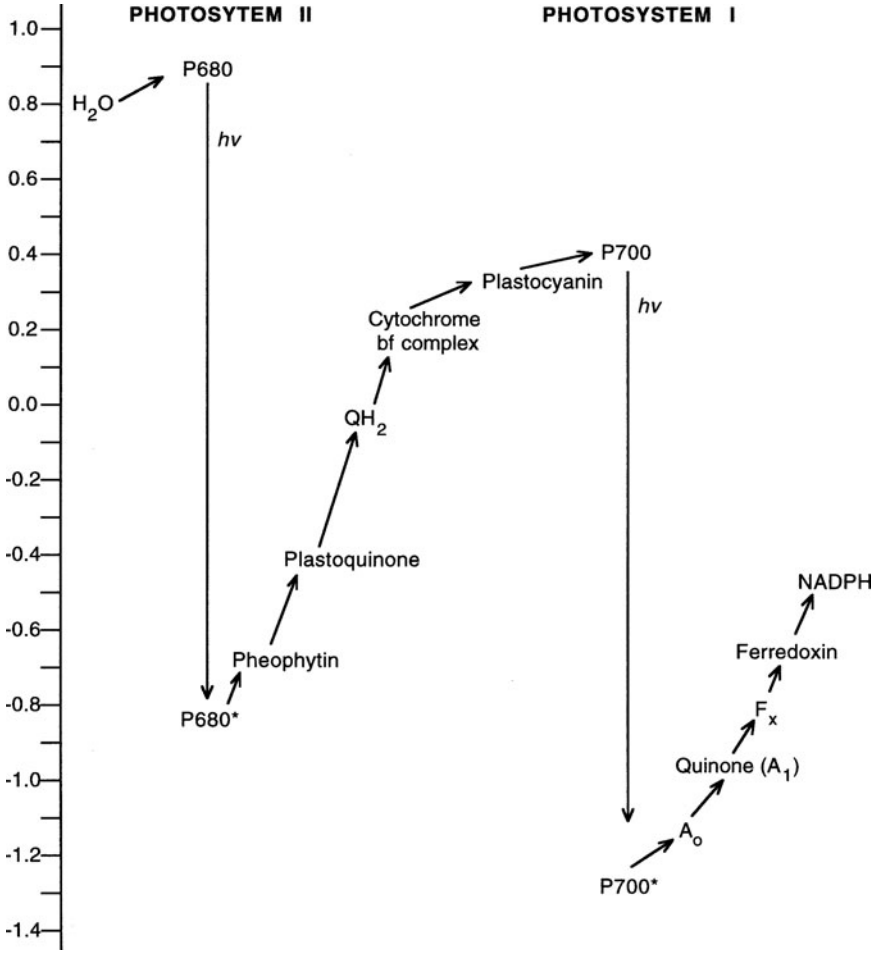


Fig. 3.5 Outline of the system of photosynthesis. The scale is the electrochemical potential (volts)

photosystem I as in photosystem II, chlorophyll molecules capture photons and shuttle the energy to a reaction center (P700) which consists of two chlorophyll *a* molecules. This reaction center becomes excited (P700*) and an electron is transferred to a monomeric acceptor chlorophyll (A_0) so that a charge-separated pair ($P700^{*+}$–A_0^-) is created. The reduction potential of A_0^- is the highest reducing potential known in biology. The electron is transferred through a series of other complexes: a quinone, an iron–sulfur complex, and finally to *ferredoxin*. The electrons in reduced ferredoxin are used to reduce NADP to NADPH via the enzyme *ferredoxin–NADP*$^+$ *reductase*. NADPH provides the reducing power necessary to fix carbon in the “dark” reactions of photosynthesis. These dark reactions are a series of reactions that couple the stored energy into metabolic chemical reactions.

If we consider the electron path in the non-photosynthetic cell, we see one of those delightful symmetries often found in scientific investigation. The biochemical pathways that lead to the oxidation of glucose are the glycolytic and citric acid cycles (Figs. 3.6 and 3.7). The electrons that are removed in these steps are used

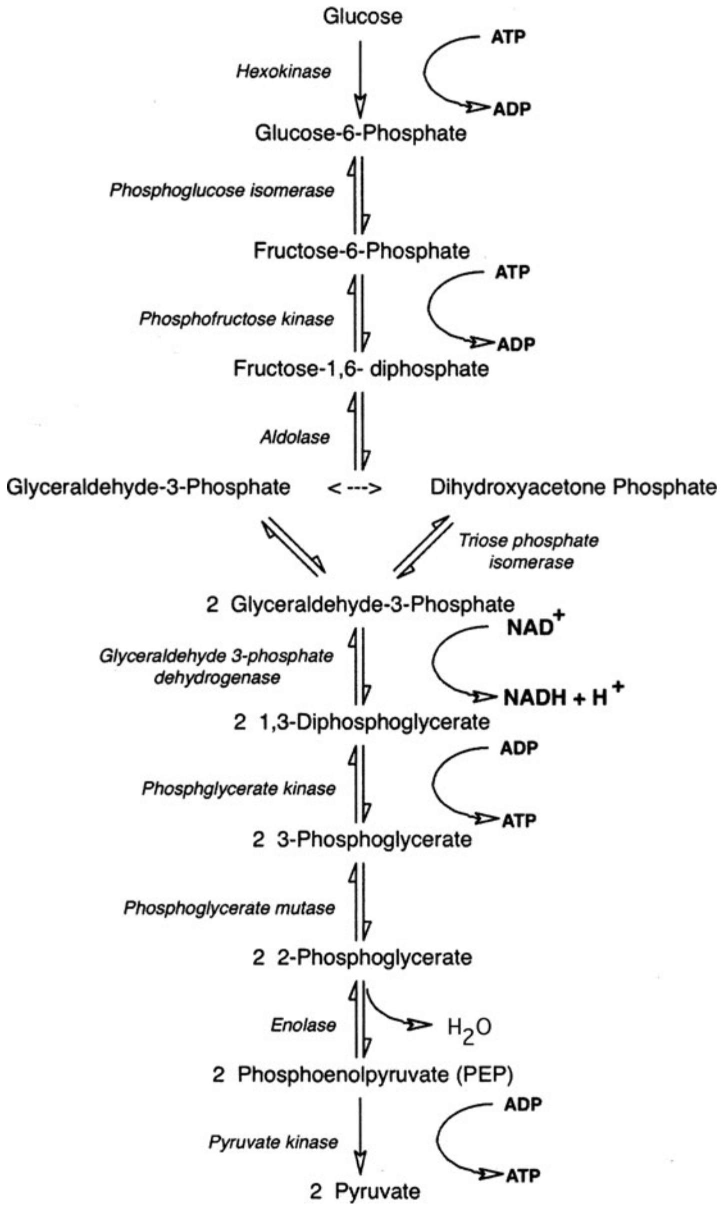


Fig. 3.6 Scheme of the glycolytic pathway (Embden–Meyerhoff)

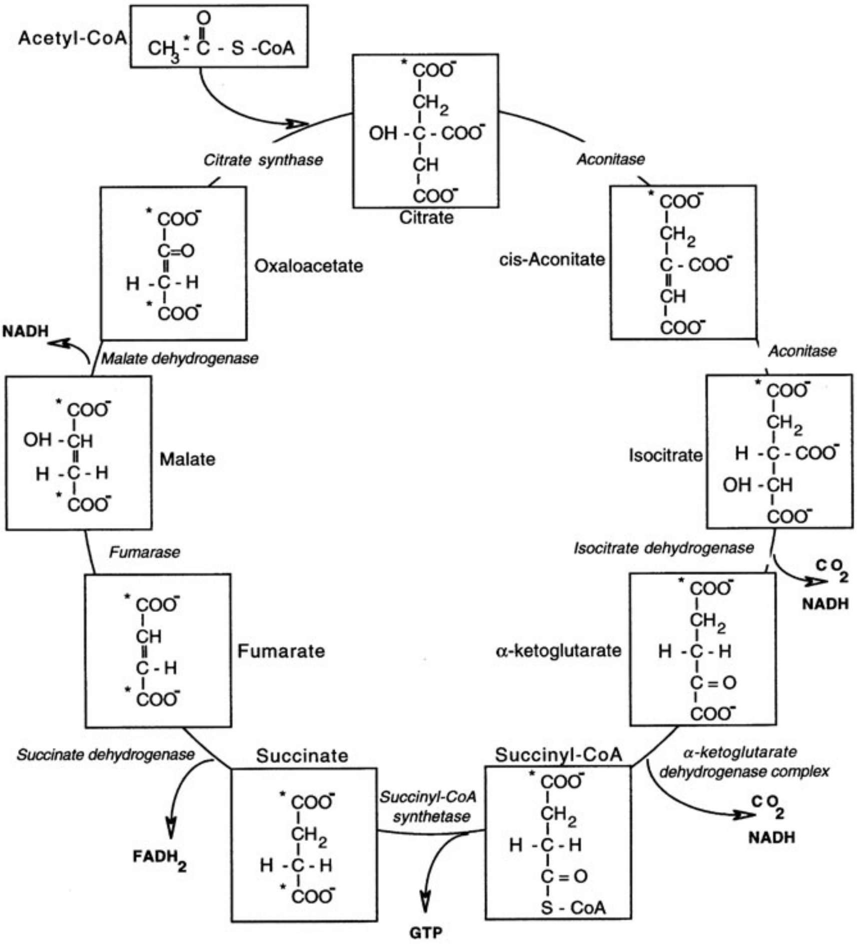


Fig. 3.7 Schematic of the citric acid cycle (Kreb's)

to reduce NAD to NADH. NADH is a transfer factor that brings electrons to the electron transport chain. Here the electrons are passed through a series of electron transfer complexes similar in many ways to the systems of photosynthesis but in reverse. The terminal electron acceptor is oxygen which is ultimately reduced to H_2O. In the process of electron transport, proton gradients are established across the mitochondrial membrane which drive the formation of ATP in fashion similar to the processes in the thylakoid membranes.

We have introduced structure into our description of biological electron transport by considering the grouping and organization of the photosystem and cytochrome complexes. Furthermore, we have implied the need for surfaces and membrane bounding as necessary for the establishment of the gradients necessary to drive ATP production. Membranes are not just separators and compartmentalizers but are part of the necessary structure of the cell machinery. Entire books are written on the structure of single elements of the photosynthetic and electron transport chains and we will examine certain details as examples in later chapters. This consideration that structural elements may form boundaries for certain subsystems and at the same time be an essential element of a related subsystem accounts for some of the complexity of biological systems. The discipline of biophysical chemistry is useful in the elucidation and understanding of such complexity. We now turn our attention to certain aspects of biological complexity.

3.5 The Cell Is a Building Block of Chemical and Biological Organization and Also a Key to the Study of Biological Complexity

Awareness of the existence of biological cells and an appreciation of cellular dimension are quite recent events in the history of science. It was in 1655 that Hooke used a compound microscope to examine sections of cork and described the small regular pores that he observed as "cells." Subsequently, Leeuwenhoek, who is credited with making the use of the microscope practical, was first to observe the single-celled organisms called *protozoa* in 1655 and then *bacteria* in 1660. The awareness of the existence of substructures such as the nucleus had to wait until improvements in technology could increase the resolving power of the early microscopes; the nucleus was not described until 1833 when Brown was carefully studying the orchid. In 1835, Schleiden and Schwann proposed their cell theory, which is perhaps the fundamental principle of the science of cell biology, namely that the nucleated cell is the basic unit from which all plants and animals are built.

In the biological sciences then, the nucleated or *eukaryotic* cell (*eu* = true, *karyon* = nucleus) is the reference point for developing a generalized abstraction about the spatial and temporal relationships of biological state space, at least for animals and plants. We know that eukaryotic biomass comprises only 50% of the total biomass in our living state space. The other 50% is *prokaryotic* organisms, the modern bacteria and cyanobacteria. These organisms are characterized by being free of a nuclear membrane and a defined nucleus; *pro* = before, *karyon* = nucleus. Although a prokaryotic cell specifically lacks an enclosed nucleus, it also lacks most of the familiar organelles such as mitochondria, endoplasmic reticulum, chloroplasts.

An advantage of biophysical chemistry is that it provides an ab initio (from the beginning) approach to understanding biological systems. If we consider the equations of state for a biological state space to be the linkages found by our ab initio approach we can build to prokaryotic cells, then eukaryotic cells and finally higher organisms by an *aufbau* (building up) process. As we build up to more complex systems, we will find that the complexity that is perceived is not because of new rules that specially govern living systems but rather because of the non-linear nature of the changes that biological evolution has undergone. Thus, the evolution of the modern biological organism may be considered as a bifurcating equation of state in biological state space lending an air of complexity and surprise that is likely to be better understood when regarded by the tools of model making. This is the essence of an active controversy in biological and molecular evolution, that is what form, continuous or discontinuous, the evolutionary function is.

3.6 A Brief History of Life

Our story begins approximately 4.5 billion years ago on the primordial Precambrian Earth, probably a youthful 1 billion years old at that point. About 1 billion years earlier, the masses of dust, whether from condensing gases from the original Big Bang or from clouds of solar orbiting meteorites, had begun to accrete because of gravitational forces and heavier particles had moved toward the center of the nascent Earth. The steadily contracting mass gave rise to great heat and violent fluxes of the melting elements, being less dense, erupted outward. These molten streams cooled, becoming dense again and were reclaimed by the Earth's gravitational field. Thus the mass of the youthful Earth churned like a giant cauldron. Geologists, using isotope dating techniques, estimate the oldest rocks in the Earth's crust to be about 4 billion years old suggesting that it took about 500 million years for the Earth's crust to form, distinct from its mantle and core. At first the crust was only several hundred meters thick. It grew thicker and cooled over the next 250 million years with the accumulation of a granite-like rock. Eventually the crust became cool enough to allow water to remain in the liquid state and early seas and pools of water formed.

Reasoned speculation holds that the Precambrian Earth was a tempest-riven world with abundant volcanic activity, violent electrical storms, and torrential rains of water. It had an atmosphere that contained no free oxygen but probably was composed at least partly from ammonia and methane. With no diatomic oxygen present, radiation shielding ozone could not have been present and the planet was bathed in an intense sheen of hard ultraviolet radiation. There were ample quantities of simple molecules such as H_2O, CH_4, CO_2, NH_3, and H_2. Given the chemistry of these molecules under conditions of heat, ultraviolet radiation, and electrical discharge, we can be reasonably confident that a high concentration of reactive species existed both in the primordial atmosphere and in the puddles and pools of collecting and evaporating water on the surface of the Earth. It has been elegantly demonstrated in the laboratory that a liquid–vapor mixture of H_2O, CH_4, CO_2, NH_3, and H_2 when exposed to conditions of heat, electrical discharge, and ultraviolet radiation will generate substantial quantities of small organic molecules including bio-organic species such as acetic and lactic acid, glycine, alanine, and aspartic acid, and urea. Thus it seems likely that the Precambrian conditions lead to a "spontaneous" generation of molecules that would be used to build biological systems.

These spontaneously formed organic species would have been concentrated in the pools of water that were constantly refluxing due to the Earth's heat and atmospheric conditions. The mud forming the floors of the pools probably contained *zeolites* which are complex silicate minerals made up of tetrahedral SiO_4 and AlO_4 groups arranged in complex three-dimensional networks. The channels in zeolites are able to adsorb ions, water, and organic molecules and hold them in a constrained geometry. Studies have shown that polymers of bio-organic molecules will form under favorable physical conditions such as heating dry organic molecules or after binding to zeolite-like minerals. It is conceivable that such a process led to early biopolymer formation. Once biopolymers form, they can to varying degree catalyze the continued formation of similar polymeric structures. Polynucleotides possess this property for self-replication. Recent studies have shown that polynucleotides also can play other catalytic roles previously reserved for polypeptides. Although proteins do not autoregenerate to the degree that polynucleotides do, protein structures such as the prion proteins may be able to self-replicate. As these early biopolymers replicated and performed catalytic functions within the confined spaces of rocks and crevices, certain molecules would have developed and adapted so that they could survive changes in the local physical environment, such as dehydration, roasting, and being used as raw materials in a competing reaction. It is likely that the natural capacity for self-replication by polynucleotides and the versatility of proteins as enzymes and building materials led to the development of a primitive chemical organization.

Most of the forces driving this natural selection of the early biomolecules would have favored those molecular communities that established ever greater control over their external physical milieu. Constructing a system in which external forces are controlled has at least two broad solutions and thus represents a bifurcation in the evolutionary equation of state. On one hand a biomolecule could manipulate its state space by becoming more complex and being able to repair, engineer, protect itself and its local symbiotic biomolecules. Such a solution is actually quite expensive if considered in energetic terms. A far more conservative and prudent solution is to build a barrier and enclose the essential control structures inside this barrier. Physical transport can be adequately maintained by diffusion and conduction. Required concentrations of chemical components can be kept high by controlling the volume in which the reactions proceed at great cost savings. The physical barrier prevents invasion of the biomolecule by other enzymes looking for substrate and reduces the chance of infection and alteration of the self-replicating genetic code. On a cost–benefit basis, therefore, the evolutionary pressure was likely to develop toward establishing a controlled internal milieu by building a barrier. We will learn how a bilipid membrane can form spontaneously when amphipathic molecules are mixed with an aqueous phase. It again seems reasonable to speculate that an external limiting membrane (probably comprised of phospholipids) provided the physical barrier bifurcation that led to the physical predecessor of the biological cell.

3.7 Evolution Can Be Modeled as a Dynamic Process with Many Bifurcations in the State Space of Life

Therefore we can consider the formation of the primitive prokaryotic cell as the first major bifurcation leading to life as we know it today. So successful was the physical barrier solution to the internal milieu problem that the prokaryotic cell, once formed, rapidly dominated the biological state space for the next 1.5 billion years. Prokaryotic cells have a single external limiting bilipid membrane often coated with a rigid cell wall consisting of carbohydrate, no internal organelles, a simple nucleotide genetic string, and a limited number of polypeptide tools and small molecules. From this description comes the classic view of the cell as a bag of water containing a variety of solutes. Prokaryotic cells are quite small and are capable of rapid replication by binary fission, i.e., dividing in two. Under conditions of surplus food (supplying energy and building materials), bacteria can replicate every 20 min. At such a rate, a single bacteria will undergo 36 divisions in 12 h and can give rise to nearly 1×10^{12} cells. This growth rate gives prokaryotic cells a tremendous advantage for adaptation. Bacteria can be found to evolve within weeks to become resistant to antibacterial agents or to metabolize a novel carbon substrate. The early prokaryotes were unlikely to have had sophisticated metabolic pathways and in general evolved to take advantage of the chemical species readily available in the local external milieu. Initially the metabolic pathways for generating energy and building materials would have been anaerobic since the physical environment of the Earth was strongly reduced (H_2O, CH_4, NH_3, and H_2). Early prokaryotes were able to live in a wide variety of extreme conditions (Table 3.8).

Table 3.8 Environmental conditions under which modern prokaryotic cells thrive

	Temperature	pH	Salinity
Psychrophiles (pseudomonads)	0°C (cold waters)		
Thermophiles (bacilli)	90°C (hot springs)		
Acetobacter (vinegar formers)		3	
Sulfur oxidizing (H_2SO_4 formers)		0	
Urea splitters		9	
Halophiles			30%

Although we cannot know the details of the primordial Earth, the diverse environmental ruggedness of modern bacteria provides an intriguing framework for speculation

3.7.1 The Scarcity of Energy and Chemical Resources Is a Fundamental Challenge Encountered in Biological Evolution

The initial chemical plenty that surrounded the early prokaryotes could not be sustained and new energy and material sources had to be used if a cell and its progeny were to survive. An important source of raw material for biopolymers was the carbon and nitrogen tied up in atmospheric CO_2 and N_2. In these forms carbon and nitrogen are fully oxidized and so, while plentiful, are extremely stable and both kinetically and thermodynamically unavailable. Carbon from CO_2 can be made available for synthetic pathways by reduction to an aldehyde, alcohol, or hydrocarbon but only through an energy-dependent reaction in which a strong reducing agent such as $NADPH_2$ adds electrons to the carbon atom. This process is called *carbon fixation*. To generate a strong reducing agent such as $NADPH_2$ requires a source of electrons and an energy source to "cock the gun." *Photosynthesis* is the mechanism by which the ancestral green-sulfur bacteria obtained the energy from light to cause the photoreduction of $NADPH_2$. The source of electrons was geochemically produced H_2S. As the electrons from H_2S were shuttled to the $NADPH_2$-like molecule, the oxidized S was released. The electrochemical constraints on green-sulfur bacteria are not terribly arduous since the redox potential of H_2S is –230 mV and the redox potential of $NADPH_2$ is –320 mV. A single photon of light is able to generate enough electrochemical energy to reduce $NADP^+$ to $NADPH_2$.

3.7.2 The Biochemical Solution to the Energy Limitations Created a New Waste Problem: Global Oxygenation

The most abundant electron donor, however, was H_2O. Ideally this source of electrons could provide virtually unlimited reducing equivalents for carbon fixation. However, H_2O has a redox potential of +820 mV thus requiring a significantly larger amount of energy to transfer electrons from it to $NADP^+$. The cyanobacteria, which developed nearly 3×10^9 years ago, solved the problem of water splitting as a source for reducing equivalents by combining a photosystem derived from green bacteria with one derived from purple bacteria. This chain of photosystems (II and I) allowed the enormous potential barrier between water and $NADP^+$ to be successfully bridged. Stripping electrons from H_2O for carbon fixation made two major bifurcations in biological state space: (1) it released biological organisms from the requirement of remaining physically near sources of H_2S such as geothermal vents and because water was omnipresent, biological life could spread without restraint across the planet's surface. (2) The by-product of splitting water was free diatomic oxygen. As the photosynthetic bacteria spread into an ever widening array of ecological niches, their constant release of oxygen into the atmosphere changed its composition to one of an aerobic environment. We know that the oxygen content of the atmosphere at first was buffered by the rusting of the Earth as ferrous iron was oxidized to ferric oxides. The bands of mineral rust in the Earth's crust date from about 2.7×10^9 years ago. Once most of the minerals were oxidized, O_2 began to accumulate in the atmosphere. The presence of O_2 in the atmosphere leads to the development of aerobic prokaryotes which were able to utilize oxygen as the terminal electron acceptor as they biochemically burned (reduced or fixed) carbon atoms. The aerobic prokaryotes constructed an electron transport chain that extracted the energy used by the photosynthetic organisms to construct new organic molecules releasing as by-products H_2O and CO_2. In so doing the aerobic prokaryotes completed stable carbon and oxygen cycles that were powered by the Sun. So naturally symmetrical are these cycles that modern purple bacteria can switch easily between photosynthesis and aerobic respiration depending on the available light or O_2. This was the beginning of a global ecology.

The cytoplasmic membrane of prokaryotes is the site of osmoregulation, transport and exchange of materials, cell respiration, electron transport, and cell wall synthesis. It plays a role in the replication and separation of the chromosomes and in cell division. The intracellular space contains only an amorphous matrix comprised of ribosomes, DNA freely floating as a fibrillar network and granules that contain food or energy storage compounds (glycogen, β-hydroxybutyrate, inorganic metaphosphate, and starch). Since so much of the metabolic activity of the prokaryotic cell occurs at the site of the plasma membrane these organisms are best thought of as heterogeneous chemical systems. The membrane not only defines the internal/external milieu bifurcation but also defines a homogeneous/heterogeneous chemistry bifurcation. With such significant metabolic support from the surface of the cell, a new problem arises: there is a limit to the volume and size supported by such cells. Invaginating the surface membranes is an efficient method of increasing the surface area-to-volume ratio of a cell without making the cell too large to be sustained by the constitutive biochemical elements. Many prokaryotes, especially the gram positive bacteria have intracellular invagination of the plasma membrane called *mesosomes* which increase the surface-to-volume ratio without massively changing the size of the cell.

3.7.3 The Response to the New Biochemical Environment Resulted in a Biological Bifurcation: The Appearance of the Eukaryotic Cell

The oxygen-rich atmosphere induced a major stress on the biological organisms that could not tolerate the highly reactive oxygen molecules. Many cells developed protective mechanisms to defend against oxygen’s highly reactive metabolites which could easily and randomly damage virtually any of the biomolecules on which life depended. One method of coping with an aerobic environment by a previously anaerobic cell is to incorporate an aerobic cell into the anaerobic one as an “independent contractor” who supplies energy to the remainder of the cell. This hypothesized symbiotic relationship between a large anaerobic cell and an aerobic cell, which then becomes the precursor to the modern mitochondrion, is called the *endosymbiotic hypothesis*. Substantial genetic evidence supports this hypothesis but it seems that the acquisition of mitochondria and chloroplasts by endosymbiosis is only one small part of a significant bifurcation in state space. This bifurcation is the development of independent intracellular organelles that characterize the eukaryotic cell (Fig. 3.8). More generally, after 2 billion years, the change from prokaryotic cell to eukaryotic *system* is a bifurcation that implies a surprising new equation of state: we have moved beyond homogeneous/heterogeneous chemistry to non-homogeneous physical systems; the eukaryotic cell is a living element comprised of multiple specialized

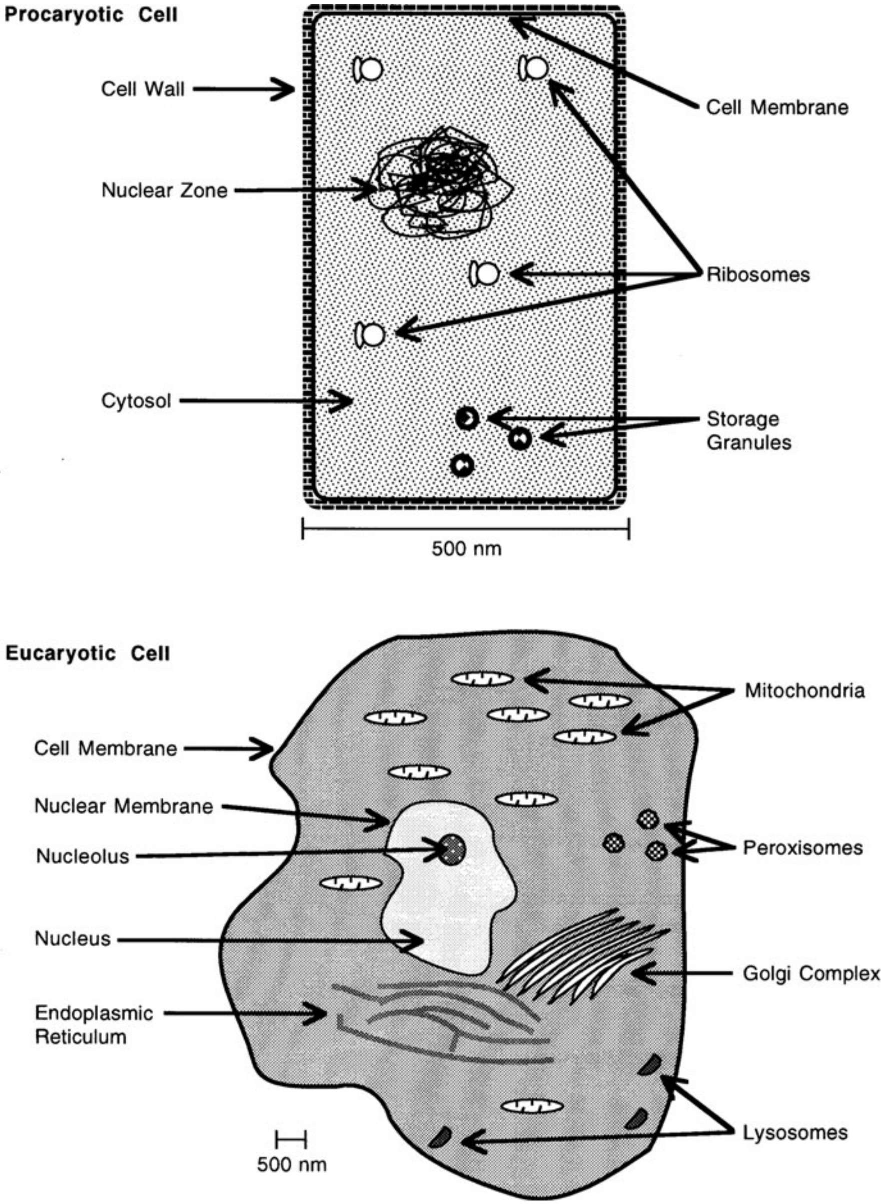


Fig. 3.8 Comparison between a typical prokaryotic and eukaryotic cell

elements that function as an integrated system. Such a system can be used as a subsystem in a recursive physiological system that will be used over and over from the organ to organism to social organizational structure. Adaptation becomes a matter of rearranging the subsystems and elements in an infinite variety of patterns. The eukaryotic system which we abstractly call a cell is the ultimate accomplishment of evolution. It is just as Schleiden and Schwann proposed: the eukaryotic cell is the essential building block for all forms of higher life. Ultimately it is the cybernetics of the system that matters because every element of each subsystem has the capacity to form every other part of the entire whole.

Thus the common view that cells are bags of salt and water is a simplistic but acceptable abstraction when considering prokaryotic cells. However, it is an inappropriate abstraction when discussing eukaryotic cells because such an abstraction eliminates the fundamental difference between eukaryotes and prokaryotes! Eukaryotic cells are distinguished by their internal membranes and membrane-bound organelles that serve to compartmentalize the wide variety of functions necessary to make a complex organism. It is true that adding internal membranes is an efficient method of increasing the surface area-to-volume ratio of a cell without making the cell too large to be sustained by the constitutive biochemical elements but this argument misses the (bifurcation) point.

The organelle as cell-membrane argument appears to be an effort to maintain the idea that the internal milieu of the cell can be considered the same as that of a prokaryotic cell. But the essential observable that differentiates a eukaryotic state space from a prokaryotic state space, the existence of internal membranes and organellar compartmentalization, is now trivialized. This argument proposes a simple linear transformation of the prokaryotic cell membrane into the internal coordinate space of the eukaryotic cell. By this argument there is no essential difference between a prokaryotic and a eukaryotic cell! This suggests that the whole of cellular biology, physiology, and anatomy is but a minor extension of prokaryotic state space. It does not seem reasonable to expect such an abstraction to survive careful scrutiny.

Why did the eukaryotic catastrophe occur? It is seems likely that as the environment continued to change and survival depended on ever increasing competition for limited resources and space, evolutionary pressures demanded an adaptability that included the ability to manipulate the environment. Once the eukaryotic cell as a system was established, the organism could carry its own internal milieu and travel, hunt, cultivate, herd, and alter its external milieu to the needs and pleasures of the organism itself. In a competitive world, such tools seem minimal for continued survival in all of the available niches. The key element in the eukaryotic catastrophe is the central role played by the genetic code and its tremendous complexity. Since eukaryotic systems are dependent on cybernetics it is only logical that the control surfaces are placed in the central position. The *chromosomes* are enclosed in the *nuclear membrane*, protected with elaborate machinery (*histones*). Careful systems for encoding and decoding and error checking must be developed to ensure the accurate and efficient transfer of information. Informational variety (informantics) is crucial for survival; and so replication is ultimately coupled with sexual reproduction to optimize the broadest possible acquisition of information.

3.7.4 Compartmentalization Is an Important Reordering of Physiochemical Relationships that Changes the Physical Environment from Solution Dominated to Surface Dominated

As previously emphasized, the idea that a cell is a lipid enclosed bag of salt and water misses one of the essential features of biological organization. There are literally thousands of interconnected biosynthetic pathways going on in the cell at any one time. In general many of these reactions can be regarded as occurring independently in the relatively constrained space of the cell. While they are organizationally linked at specific control points they do not act like reactions in a homogeneous or even heterogeneous chemical system. The relative isolation of these disparate chemical reactions until a supraorganization structure integrates specific reactions is accomplished through the use of *compartmentalization*. In the cell, the principal mechanism of compartmentalization is the membrane-enclosed *organelle*. General control mechanisms are in the domain of biophysical chemical processes, but many of the details of organellar, cellular, and organismal control are the domain of physiology. We will concern ourselves with the rules for the overall system which are defined by biophysical considerations. The savvy student will see application of these considerations throughout his or her study of biology. Fundamental to the examination of the cell in terms of its physical organizing principles is an awareness of the spatial dimension of the cell and its organelles. This includes an appreciation of the size of the constituent components making up a cellular system, the atomic, molecular, solvent, macromolecular, and supramolecular organizations (Table 3.9).

On a molar basis, water is the most prevalent chemical species found in biological systems. This has led to the reasonable assumption that water is the predominant solvent in biological systems. Water is, however, a unique solvent by virtue of its extensive hydrogen-bonded character in the bulk phase. If the description of the biological cell as a bag of water filled with a dilute aqueous solution of biomolecules were correct, the bulk properties of water would likely be very important. However, the cell is not a bag of homogeneous aqueous solution but rather a complex array of membranes, colloidal organelles, and a bewildering array of densely packed macromolecules having the character more of a hydrophilic gel than of a homogeneous solution of water. This combination means that the cell is a heterogeneous, multi-phasic, surface-dominated environment rather than a homogeneous solution of many species. The true role of water in this environment cannot be easily predicted since the nature of the chemical reactions will depend to a large extent on the *non-homogeneous* behavior of cells. The term non-homogeneous was coined

Table 3.9 Size of biological components

Biological structure	Dimension	Shape
Plant cells	50–100 μm	Polyhedral
Animal cells	10–70 μm	Variable
Platelet	3 μm	Discoid
Red cell (erythrocyte)	7 μm	Discoid
White blood cell	12 μm	Round
Muscle cell (skeletal)	30 cm (length) 10–100 μm (diameter)	Cylindrical
Muscle cell (smooth)	30–200 μm (length) 5–10 μm (diameter)	Spindle
Liver cell (hepatocyte)	25 μm	Polyhedral
Nerve cell (neuron)		
Body	5–150 μm	Variable
Axon	Up to 100 cm long	
Bacteria	1–5 μm	
Virus particle	20–50 nm	
Ribosome	20 nm	
Amyloid fibril diameter	10 nm	
Hemoglobin	65 Å	
Myoglobin	36 Å	
Glucose	7 Å	
Amino acid – alanine	5 Å	
Water molecule	2.4 Å	
Water, tetrahedral cluster	5 Å	
Sodium ion (crystal radius)	0.95 Å	
Sodium ion (hydrated)	5 Å	
Potassium ion (crystal radius)	1.33 Å	
Sodium channel inside dimension	5 Å	
Potassium channel inside dimension	3 Å	
Double layer dimension	5–10 Å	
Lipid bilayer thickness	40–50 Å	
Mitochondria	0.5 μm × 1.5 μm	
Lysosome		
Primary	50 nm	
Secondary	200–500 nm	
Nuclear pore complex	50 nm	
Coated pit (clathrin coated)	150 nm	
Nucleus	3–10 μm	
Diameter of DNA double helix	20 Å	

by Freeman for chemical processes that cannot occur randomly in space because at least one component is not randomly distributed in either a single or a micro-heterogeneous phase. In comparison to *homogeneous processes*, where components are randomly distributed in a single phase, and macroscopically *heterogeneous processes*, where components from one or more phases interact at the interface between the phases, non-homogeneous processes may well describe the biological system most accurately.

Further Reading

General Reading

Most readers will already have one or more textbooks that will explore biology, cell biology, and biochemistry. Among those I have found useful are

Alberts B., Bray D., Lewis J., Ruff M., Roberts K., and Watson J. (2007) *The Molecular Biology of the Cell*, 5th edition. Garland Press, New York.

Alon U. (2007) *An Introduction to Systems Biology: Design Principles of Biological Circuits.* Chapman and Hall/CRC, Boca Raton, FL.

Berg J.M., Tymoczko J.L., and Stryer L. (2007) *Biochemistry*, 6th edition. W.H. Freeman, New York.

Nelson, D.L. and Cox M.M. (2009) *Lehninger Principles of Biochemistry*, 5th edition. W.H. Freeman, New York.

Molecular Evolution

Allègre C.J. and Schneider S.H. (1994) The evolution of the earth, *Sci. Am.*, **271, 4**:66–75.

Cech T.R. (1986) RNA as an enzyme, *Sci. Am.*, **255**, **5**:64–75.

Miller S.L. (1987) Which organic compounds could have occurred on the prebiotic earth, *Cold Spring Harbor Symp. Quant. Biol.*, **52**:17–27.

Orgel L.E. (1994) The origin of life on the earth, *Sci. Am.*, **271, 4**:76–822.

Rebek J. (1994) Synthetic self-replicating molecules, *Sci. Am.*, **271, 1**:48–55.

Biological Evolution

Alberts B. (1998) The cell as a collection of protein machines: Preparing the next generation of molecular biologists. *Cell*, **92**: 291–294.

De Duve C. (1996) The birth of complex cells, *Sci. Am.*, **274, 4**:50–57.

Fredrickson J.K. and Onstott T.C. (1996) Microbes deep inside the earth, *Sci. Am.*, **275, 4**:68–73.

Gould S.J. (1994) The evolution of life on the earth, *Sci. Am.*, **271, 4**:84–91.

Kauffman S.A. (1991) Antichaos and adaptation, *Sci. Am.*, **265, 2**:78–84.

Knoll A.H. (1991) End of the proterozoic eon, *Sci. Am.*, **265, 4**:64–73.

Margulis L. (1970) *Origin of Eukaryotic Cells*. Yale University Press, New Haven, CT.

Bioenergetics

For a readable survey of the components in metabolic and photosynthetic pathways, see:

Anselme J.-P. (1997) Understanding oxidation-reduction in organic chemistry, *J. Chem. Educ.*, **74**:69–722.

Brudvig G.W., Beck W.F., and de Paula J.C. (1989) Mechanism of photosynthetic water oxidation, *Ann. Rev. Biophys. Biophys. Chem.*, **18**:25–46.

Govindjee and Coleman W.J. (1990) How plants make oxygen, *Sci. Am.*, **262, 2**:50–58.

Youvan D.C. and Marrs B.L. (1987) Molecular mechanisms of photosynthesis, *Sci. Am.*, **256, 6**: 42–48.

Regulation of Processes in Cell Biology

The organization of the cell is discussed in

De Camilli P., Emr S.D., McPherson P.S., and Novick P. (1996) Phosphoinositides as regulators in membrane traffic, *Science*, **271**:1533–1539.

Fuchs E. and Weber K. (1994) Intermediate filaments, *Annu. Rev. Biochem.*, **63**:345–382.

Görlich D. and Mattaj I.W. (1996) Nucleocytoplasmic transport, *Science*, **271**:1513–1518.

Grunstein M. (1992) Histones as regulators of genes, *Sci. Am.*, **267, 4**:68–74.

Rothman J.E. and Orci L. (1996) Budding vesicles in living cells, *Sci. Am.*, **274, 3**:70–75.

Schatz G. and Dobberstein B. (1996) Common principles of protein translocation across membranes, *Science*, **271**:1519–1526.

Scherman R. and Orci L. (1996) Coat proteins and vesicle budding, *Science*, **271**:1526–15322.

Vallee R.B. and Sheetz M.P. (1996) Targeting of motor proteins, *Science*, **271**:1539–1544.

Biological Energy Transduction

Cramer W.A. and Knaff D.B. (1991) *Energy Transduction in Biological Membranes. A Textbook of Bioenergetics.* Springer-Verlag, New York.

de Silva A.P., Gunnlaugsson T., and McCoy C.P. (1997) Photoionic supermolecules: mobilizing the charge and light brigades, *J. Chem. Educ.*, **74**:53–58.

Scherer S. (1990) Do photosynthetic and respiratory electron transport chains share redox proteins? *TIBS*, **15**:458–462.

Youvan D.C. and Marrs B.L. (1987) Molecular mechanisms of photosynthesis, *Sci. Am.*, **256, 6**: 42–48.

Problem Sets

1. A prokaryotic cell has an energy requirement of 100 μmol of substrate per division. It will divide every 40 min as long as it has excess substrate available. How many bacteria will be present in the following situations?

   (a) 5 M substrate, no time limit.
   (b) 5 M substrate, 10 divisions.
   (c) 1 M substrate, no time limit.
   (d) 1 M substrate, 10 divisions.

2. The concept of the cell as a container of dilute aqueous solution can be supported only if certain abstractions are made. List some of these.
3. Consider the lipid membranes in the cell. Which organellar membranes are equivalent?
4. Given the internal size of a mitochondrion, how many water molecules could be contained within one? What is the molarity of this compartment?
5. Does the endosymbiotic theory support the view that compartmentalization is causally related to (a) the surface–volume problem or (b) the oxygenation–energy catastrophe?

Chapter 4
Physical Thoughts, Biological Systems – The Application of Modeling Principles to Understanding Biological Systems

Contents

4.1 The Interaction Between Formal Models and Natural Systems Is the Essence of Physical and Biophysical Science

In the chapter two, we codified the scientific method as a sequential series of models that first organize description, then explanation, and finally experimental verification in a coherent manner. The accurate and effective coding or mapping of a real system (and its description) onto a formal/mathematically cogent system that correctly serves as an abstract model of the real world is the task of the field of systems science. The progression of inquiry is a knowledge operator that supervises

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_4,

the scientific method such that a systems science modeling step is incorporated at each inquiry step. The ideal outcome of this essential scientific enterprise is that the formal system and its behavior can be compared and found to be invariant with the behavior of the real system that it represents. This last requirement is the experimental essence of systems science itself.

The use of formal models to describe a natural system has been a central principle in science since the pre-Socratic Greeks (see Appendix C for a historical perspective). The critical use of numerical experimental data to refine and validate a formal mathematical model is the contribution of Galileo and Kepler. We are interested in a description of a natural system that allows a formal model to be constructed, which, for a precisely defined subsystem, is invariant with the natural system. The following questions about the progression of inquiry and its relationship to model-making capture most of the basic issues in the philosophy of science – an epistemology of scientific modeling. Although no complete answers will be forthcoming (or can reasonably ever be expected) we will consider several central points.

First of all, what is there about the nature of natural systems that requires modeling at all? Why not just describe the thing fully and be done with it? What after all is a model and what constitutes a "good" model?

Assuming that we need models for some useful purpose, how do we build them? What is the nature of the observable quality and quantity and how do we map it from the natural system, N, onto the formal system, F?

If we are compelled to go through this process once, how can we prevent having to do it again unless absolutely necessary? In other words, can we find simplifying rules that make a model more straightforward to construct and more easily understood? How can we recognize systems and models that are similar thus allowing us to generalize and solve any particular problem just once. Can we use the rules of identifying similar systems to justify our use of our simplified or abstracted models? If F_1 and F_2 are similar does this imply that N_1 and N_2 are also similar?

Finally if these models are general enough and the observables can be linked to one another in a dependable manner, can we generalize further, and, based on our model, find and describe "laws of nature"?

The central role of constructing models in scientific investigation should be visible at his point.

4.2 Observables Are the Link Between Observer and Reality

Study and knowledge of natural systems in a scientific manner depends on the assignment of specific observables that describe a system and characterize the linkages between the observables. Although there is room for theory and speculation in the process of investigating a system, contact with reality is made through the observables. We construct models of a natural system for the purpose of bringing order to our observations and experiences about the natural system of which we are a part. This sense of order enables us to make specific predictions about the world of our experience. Building a model then is consistent with the earliest scientific philosophy that had its origins with the Ionian philosophers. A model must be credible in terms of the observables and their linkages found in the natural system. A model of a natural system, N, is essentially a generalization from study of a subsystem of N, so that to a specified degree of accuracy, the resulting model reflects the original system. How well the model performs this task determines the goodness of the model. Because a model is built from consideration of a subsystem of the more general natural system, the ability of a model to make good predictions about the future state of the natural system depends to a large extent on the relationship of the subsystem as an effective and useful window to the more general natural system. The best any model can do is to capture the subsystem under study and so its “goodness” is correlated to the system:subsystem relationship. It follows that the goodness of a model depends on how the natural system is characterized by the chosen observables, the choice of the observables used to describe the subsystem, and the manner by which the observables are mapped or encoded into a formal mathematical system that then represents the system under consideration. Thus the formal mathematical system, F, is the model that we manipulate and use.

Consider the most familiar simple system known to students of chemistry, an ideal gas enclosed in a container. To describe this gas, we *choose* as our observables: its temperature, pressure, and the volume it occupies (P, V, T). With these observables, the actual physical states of the system that give rise to P, V, T can be mapped into a three-dimensional space, R^3, whose coordinates are P, V, T. The states described by P, V, T are *abstract states* that represent the actual physical states of the system. This means that we know that the representation of the actual physical nature of a gas by the three chosen observables is an abstraction of the natural system because *we have explicitly chosen to ignore all the other observables* in the system that also will influence the gas. We also know that our chosen observables are not independent at equilibrium but are linked by an *equation of state*, the familiar, $PV = \alpha T$ where α is an appropriate proportionality constant. We also have learned before (and will cover this point again in subsequent chapters) that the macroscopic behavior of an ideal gas is well predicted if our three observables are known and linked by the equation of state. All ideal gases will map faithfully onto the abstract space R^3, because we have defined an ideal gas as being wholly and accurately described by only these three observables. It turns out that at very low pressures most real gases, for which R^3 is truly an abstract state space, will also be reasonably well represented by the abstract state space. In spite of the close correspondence deviations from the abstract state space will be found experimentally because the abstraction process eliminated from consideration observables that need inclusion and linkage in the equation of state. Even for an ideal gas, the observables and equation of state are useful only for describing a macroscopic system and, if we are interested in the predicted behavior of this gas, at microscopic level, then the observables and the state space will change. Remember, however, that we are still describing the same natural system.

If we consider the gas as composed of a mole of molecules that are moving about, we might choose to represent each molecule mechanically and describe its position and momentum. These observables would be mapped into a space, $R^6 \times 6.02 \times 10^{23}$, and the linkages between the observables would be given by the laws of mechanics.

4.3 Systems Science Guides the Linkage of Natural and Formal Models

At each stage of the progression of inquiry, the biophysical chemist explores a natural system at the levels of description, causality, and experimentation. Each of these explorations generates an abstracted representation that must be mapped onto a formal model. The formal model can be manipulated with feedback to the natural system usually in the form of predicted values that can be compared with further descriptive or experimental observations. The overall process of scientific inquiry is an array of abstracted models at a level of epistemological detail selected by the choice of observables. Each of these can be mapped to a formal model system. There will be interaction between these various models, both formal and epistemological. This overall interaction can be imagined as a modeling matrix in which the descriptive, explanatory, and experimental models that are the progression of inquiry can each be formally represented in mathematical form and symbolism – a mapping of one form of model system to another often more general form. A simplified and singular representation of the relationship between a natural system and a good formal model can be seen in Fig. 4.1 in a system called the modeling mirror. On one side is the real world, composed of the natural systems that are studied by an observer

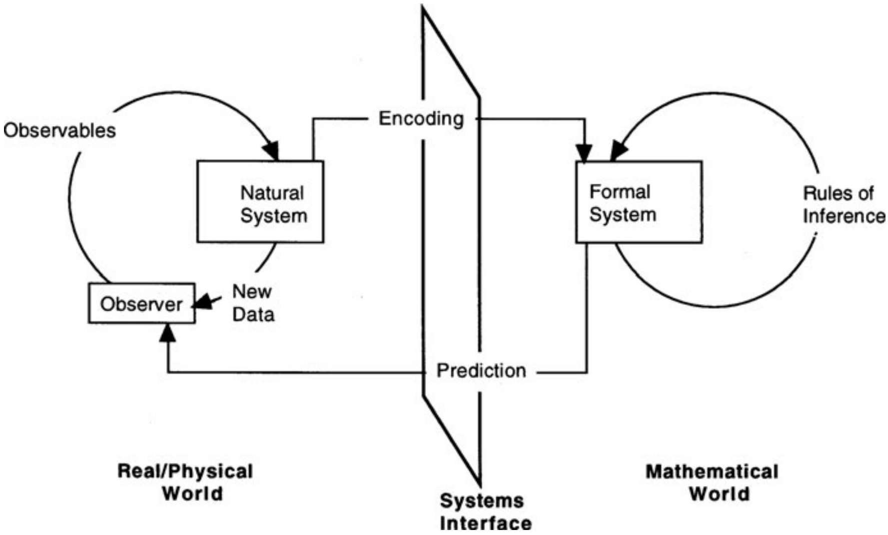


Fig. 4.1 The modeling mirror (From Casti, 1992a. Reprinted by permission of Wiley)

inevitably operating within the natural system itself, for knowledge about the system requires data derived from observation and measured by reference to a set of chosen observables. The physical scientist's job lies primarily in this world, observing and recording data in terms of the observables and using his skill to assure that the data are as complete and faithful as possible. On the other side of the mirror is the mathematical world, which itself is comprised of a variety of formal systems. We will generally be interested in formal systems of mathematical familiarity such as systems of differential equations, topological systems, and finite groups. Each of these systems has its own internal set of axioms, syntax, and logical inference, the discovery of which is the domain of the mathematician. Formal systems are the creation of the human mind and are defined in terms of symbols and the rules for symbolic manipulation; natural systems have a reality defined by actual and tangible observables.

In a sense the modeling mirror shows two images, one virtual and the other real and, if the models are well done, apparently indistinguishable. Since in model making we are interested in organizing our observations from the real world into a logical pattern from which inference can be dependably made, it is clear that the links between the two worlds seen through the mirror must be carefully formed. A fundamental task in model building is to encode correctly the observables and their linkages into a reliable correspondence with the syntax and rules of inference of the formal system. This allows us to use the rules of the formal system to derive new connections between the observables and linkages of the natural system. This has great importance to the experimental validation in modern science. If:

- the observables are properly chosen and reliably determined and,
- these (observable's) values are correctly encoded onto the formal system,
- then the formal system may be used to infer new theorems or make predictions about the natural system.

These predictions must then be decoded and transformed again into the language of observables such that the physical or natural scientist can compare the formally derived prediction with real observables. Thus the modeling mirror is a dynamic process involving the physical, mathematical, and systemic interactions. Special care must be taken in the process of encoding and decoding the natural and formal systems one onto the other. The choice of observables that properly represent the natural world and which can be usefully manipulated in a formal system is an essential step in model making.

4.4 Abstraction and Approximation May Be Useful but Are Not Always Correct

In understanding the process of developing the links between the theoretical and experimentally validated models it is important to appreciate that many of our attempts to reduce knowledge to laws and the resulting abstractions often lead to laws that are approximate. The physicist Richard Feynman made this point in the very first of his *Lectures on Physics*. Approximations to the laws of nature are inevitable because in many cases our knowledge is incomplete. What is incomplete is often our ability to measure the system accurately. In other words either the observables are measured inaccurately or they are the wrong observables for the system. Feynman's example is instructive by its simplicity: The mass of an object seems never to change; a spinning top weighs what a still top weighs. And so we write a law: mass is constant, independent of speed. Fine, except the law is wrong if we approach the speed of light. Then the mass increases. So the true law is that as long as an object is moving at less than 150,000 m/s the mass is constant to within one part in 1 million. Now if we properly constrain the system in which we are operating, we can use the approximate law that mass is independent of speed with more than adequate accuracy for our experimental purposes. Empirically we are on predictable ground. However, philosophically we are completely wrong in using the approximate law. Once we recognize that mass, for example, is not constant, our entire worldview must radically change. The approximate law becomes a special case, but one that is really little more than a parlor trick. This philosophical invariance makes it crucial for the serious student and all practitioners of science to appreciate the interaction between the real state space, the observables chosen to describe it, the experimental methodology used to explore the relationship between the observable and the state space, and the formal model used to abstractly connect them. We will see this epistemological duality many times in this book, and the astute scientist will experience numerous examples in daily life. We can know some observables quite well and they work within the framework of an approximate law. However, even though they may be useful and give an adequately practical result, the laws they may appear to support are in some degree wrong and must give way to more complete (and often more difficult) models.

4.5 The Choices Made in Observables and Measurement Influence What Can Be Known About a System

If we consider a subsystem of the observable natural world, S, we can describe a set of specific and distinct states that are called the *abstract states*, Ω, such that $\Omega=\{w_1, w_2, \ldots\}$. The number of elements w of Ω may be finite or infinite, and an observer of S may or may not be able to discern one of the abstract states from another depending on the resolution of the observables chosen. The observable states of S depend on the observer and the tools available to probe the system. The ability to distinguish one abstract state from another does not depend on any intrinsic property of the system but rather is dependent on the ability to define and identify separate observable states.

Figure 4.2 shows a simple system of a triangle with its vertices labeled A through C. The triangle can be rotated through space counterclockwise and can be

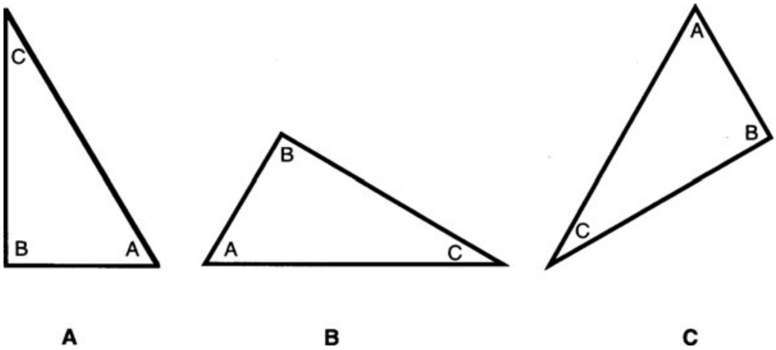


Fig. 4.2 Some of the abstract states for a system *S*

distinguished in forms that correspond to rotations of 0, $\frac{2\pi}{3}$, and $\frac{4\pi}{3}$. Three possible sets of abstract space are noted in the figure, each equally valid since each element *w* is just a representation or a label for a physical state of *S*. This example illustrates that a given system is likely to have more than one abstract state and that these states are not necessarily represented numerically. All of these abstract spaces are equivalent in their representation of the distinct states of *S* as we have drawn them because nature has no preferred space of states. The preference that a model maker may have for one abstract space Ω over another is one of convenience.

Since knowledge of the abstract state depends on the selection of observables leading to an observable state, let us see what difference the choice of measurement can make in our knowledge of a system. Assume system *S* with a set of abstract states Ω. A measurement is really a rule (or function) that allows us to map each element *w* of the state Ω to a real number. Thus the rule *f*, which does the mapping, is an observable of *S*. We can write this idea in a formal shorthand: f: $\Omega \rightarrow R$. What effect does the choice of observable or *f* have on our knowledge of the system?

We consider a rat in a maze and wish to know its position. Let $\Omega = \{w_1, w_2, w_3, w_4\}$ so that $w_1 =$ the beginning of the maze, $w_2 = \frac{1}{3}$ of the way through the maze, $w_3 = \frac{2}{3}$ through the maze, and $w_4 =$ at the end of the maze. We will define our observable, $f: \Omega \rightarrow R$ by the following rule:

$$(f\,w) = \text{distance from the beginning of the maze}$$

This rule is applied to the elements of the abstract state and:

$$f(w_1) = 0 \quad f(w_2) = \frac{1}{3} \quad f(w_3) = \frac{2}{3} \quad f(w_4) = 1$$

If we now consider another observable *g* such that g: $\Omega \rightarrow R$ as:

$$g(w) = \text{distance from the center of the maze}$$

we now get the following mapping:

$$g(w_1) = \frac{1}{2} \quad g(w_2) = \frac{1}{6} \quad g(w_3) = \frac{1}{6} \quad g(w_4) = \frac{1}{2}$$

The observable *g* is not able to separate or distinguish all of the states in contrast to *f*, which is able to distinguish them. Thus by using the observable f, a picture of the system is obtained with greater resolution (i.e., we can detail more of it). Further, *f* and *g* can be explicitly related to one another in a *linkage* relationship which formally codifies the idea that one observable is a function of the other. The linkage relationship can compensate for an observable such as *g*, which contains less information about a system than *f* if it is known. This is a vital idea because it may be very difficult to measure *f* but easy and inexpensive to measure *g*. If the linkage relationship is known, then measuring *g* ultimately gives the same information as measuring the more accurate *f*.

4.6 The Simplifying Concept of Abstraction Is Central to Both Scientific Understanding and Misconception

If we now consider the requirements, in the general case, to fully observe and describe a system, *S*, that exists in the real world, we will come to the inconvenient conclusion that we will require an infinite number of observables. The system is described by the set of abstract states Ω and the entire set of observables $F=\{f_\alpha\}$. It is usually impossible and almost certainly impractical to work with such a large number of observables; hence most observables are simply ignored or explicitly pushed to the side excluded from consideration in the system. In mathematical terms, our attention is focused on a subset, *A* of *F*. The set *A* is called an *abstraction* and represents a partial but convenient view of *S*. The usefulness of abstractions cannot be overestimated since they form one of the most practical methods for reducing the description of a system to simpler and more easily understood terms and functions (linkages). On the other hand, the amount of trouble that abstractions, especially good ones, can cause should also never be forgotten. Abstractions by their very nature can be so powerful and practical within a limited range that they rapidly become dogmatic. Instead of properly being viewed as representing a limited perception of the entire system, they may well become synonymous with a "gold standard" definition of the system (i.e., the complete description). Once an abstraction attains the status of dogma substantial efforts may be wasted attempting to force observations into compliance with the abstraction. In these cases, the abstraction itself has been mistaken for the real system. We will see repeatedly in the discussions that follow, how *modifying the assumptions* in a field of investigation, which are quite simply the abstraction of the system under consideration, becomes the linchpin to moving knowledge of the system forward.

It is in human nature to seize upon useful abstractions with an almost religious faith; the student and the practicing scientist should recognize this intrinsic tendency and counter it with a firm, frank, and skeptical knowledge of the assumptions and previously ignored observables upon which our modern theories have been constructed. The power of the scientific method that distinguishes science from magic and religion is the explicit recognition of the epistemological tool of using an abstraction to describe a system. The validity of the abstraction as a practical working model is then tested by experimentation for the purpose of connecting other observables and functions in a consistent manner to the abstraction. Viewed in this context, the scientific method is a method for critically examining assumptions and “observed knowledge.”

4.7 Equations of State Capture the System Behavior or “Systemness”

The phrase “an equation of state” is familiar when discussed in relation to the ideal gas law and to considerations of thermodynamic systems. Such discussion rarely reaches the more fundamental nature of the idea of the equation of state as reflecting the quality of “systemness” that is so crucial to a scientific systematization of a natural system.

We have seen that having a method of measuring a system N (i.e., having a set of observables $\{f_1, f_2,\ldots\}$), gives us access to knowing something about the elements that comprise N. However, just knowing each piece, even in great detail, does not provide a sense of the overall meaning or arrangement of N. Thus the ability to describe the quality of paradigmatic interaction that we intuitively feel is the “understanding” of the system is not achieved. “Systemness” then could be characterized as the observables of a system plus the relationships between observables that gives rise to an understanding of the system as a whole. This relationship between observables that gives rise to systematic understanding of N is the *equation of state* for N. The equation of state, Φ, is the set of mathematical relationships describing the dependency of the observables on one another. This can be written:

$$\Phi_i = \{f_1, f_2 \ldots, f_n\} = 0 \qquad \text{for} \quad i = 1, 2, ..., m \tag{4.1}$$

Let us consider the ideal gas law in terms of its “systemness.” We enclose our ideal gas in a container and define the set of abstract states in terms of the position and momentum of each of the particles in the gas. The pressure, volume, and temperature of the gas are the observables each measured when the gas is in state w. Our equation of state for the ideal gas relates these observables:

$$\Phi = \{P, V, T\} = 0 \tag{4.2}$$

and the form of the relationship between these is found to be

$$\Phi i = \{x, y, z\} = xy - z \tag{4.3}$$

An obvious aspect of these relationships is that once two observables are known, the third is determined by them. Thus a deterministic relationship exists between the observables defined by the equation of state. Since our “understanding” of the system is to a great degree codified by the deterministic equation of state, we often unthinkingly equate such understanding with the idea of causality. We may come to believe that one observable is caused by the others. From a formal viewpoint, an equation of state of the form given in Eq. (4.1) is deterministic because the observables all depend on one another. However, there are many equations that can link the observables to satisfy the formal relationship; hence, the equation of state cannot be considered to contain any specific information about the reasons or causes for the observed linkages. These are the core ideas that we will consider further when we discuss state and path functions in coming chapters.

The issue of *causality* is a very substantial one. On one hand, our curiosity is often directed at understanding the causal links in a system; on the other hand, we want to know how to separate observables into those that are responsible for a state from those that are epiphenomena (i.e., just accompany a change in state). Both approaches are necessary if system modeling theory is to be useful. In biological systems a common difficulty of this order can be seen in the classic nature-versus-nurture arguments. Though we may have well-developed equations of state that have been empirically derived, we may not be able to discern what is the correct causal ordering of the observables.

Let us explore this a little more deeply. In our exploration of a system, we often find a set of observables, r, which remains fixed for all of the states, w, that are elements of Ω. These so-called *parameters* are familiar to us as the acceleration of gravity, the RC constant of an oscillator, or the discount rate in the bond market. We call these observables α such that each set of $\mathrm{\alpha}_i$ describes a different system.

$$f_i(w) = \alpha_i \qquad \alpha_i = \text{a real number for } i = 1, 2, 3, \ldots r \tag{4.4}$$

We can write the equation of state in terms of the parameters, thus explicitly demonstrating that the description of the state depends on the parameters:

$$\Phi_{\alpha(i)}(f_{r+1},\ f_{r+2}, ...,\ f_{n-m}) = 0 \tag{4.5}$$

We will call this set of observables u. We have written the last term as f_{n-m} because we wish to consider a further set of observables that are functions of the set of observables just described (i.e., $f(u) = y$). The three sets of observables written in sequence are

$$a = (\alpha_1, \alpha_2, \ldots, \alpha_r) \tag{4.6}$$

$$u = (f_{r+1},\ f_{r+2}, \ldots,\ f_{n-m}) \tag{4.7}$$

$$y = (f_{n-m+1},\ f_{n-m+2}, \ldots,\ f_n) \tag{4.8}$$

We can write the equation of state

$$\Phi_\alpha (u) = y. \tag{4.9}$$

Equation (4.9) is structured so that the parameters describe the system, u, and act as inputs. The observables y are outputs to the system. Since we have written y as being solved in terms of the parameters u, a causal direction is now discernible where before the equation of state was acausal. One must observe caution here. Although the inputs may cause the outputs, it is likely that there are a variety of ways for y to be solved in terms of u, and, therefore there may be a variety of causal relationships. In physics, and to a lesser degree in chemistry, there are often adequate experiences (i.e., errors that have been discovered) to establish conventions that allow reliable sorting out of the causal order. In biology, economics, and the social sciences however, there may be no clear way of knowing which causal ordering is correct. For example, is the tendency toward violence an inherited trait, or is it a consequence of social conditions and conditioning, or perhaps of some combination of both? Even though investigators may hold strong opinions, no one really knows the causal ordering.

4.8 Equivalent Descriptions Contain the Same Information

If the descriptions of two systems each contain the same information, then the descriptions can be said to be *equivalent*. If we draw an ellipse as in Fig. 4.3 with semi-axes of length a and b and rotate the ellipse through angle θ, the figures are all different in terms of the curves drawn on the x–y plane of the paper, yet each of the ellipses is equivalent to the other as described by the analytical expression $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$. The ellipses are equivalent through the coordinate transformations, x' equals $x \cos\theta - y \sin\theta$ and $y' = x \sin\theta + y \cos\theta$. The information *about the ellipse* is identical in each case and so the descriptions are equivalent. The difference in appearance between the ellipses is a result of the way we look at each of them. The

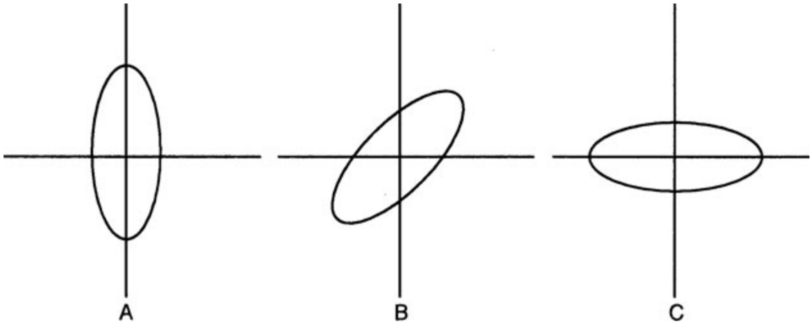


Fig. 4.3 These ellipses are equivalent because they can be shown to be identical if an appropriate coordinate transform is applied

intrinsic properties of an ellipse are given by the parameters a and b, and these are invariant regardless of the coordinate system that describes the ellipse. This important point cannot be overemphasized. Often apparent differences are coordinate dependent and thus are descriptive artifacts and are not properties of a system itself. For example, the fundamental systemic error suffered by inventors of perpetual motion machines is the assessment that a machine has the intrinsic property of producing more energy than it uses rather than recognizing that the "extra" energy was an input missed because of a coordinate-dependent error.

The ellipses just described are all equivalent because the parameters a and b are constant. However, a and b can be changed and the ellipses can still be equivalent so long as x and y are appropriately transformed as well. The transformation depends on the ratio of change of the axes (i.e., x' equals $\frac{a}{a'}$ x and $y' = \frac{b}{b'}$ y). At certain points (when $a, a', b, and\ b'$ equals 0) the transformations will clearly no longer yield an ellipse and a new type of conic section results. Such points in a model are called *bifurcation points* and are important in determining stability and transitions from one family of equivalent forms to another family of "different forms."

Descriptions of one system are equivalent to the original system if the inputs and outputs from transformed equations of state are indistinguishable. Often the coordinate changes required will be non-linear. Figure 4.4 demonstrates this equivalence in terms of a commutativity diagram. If two descriptions cannot be made the same by coordinate transformation then they will not be found to commute. D'Arcy Thompson used the principles of non-linear transformation of the physical state space to propose a theory of evolution in which transformations from one species to another occurred due to major transformations involving the whole organism rather than just successive minor changes in local body parts. An example of these types of transformations is shown in Fig. 4.5.

Fig. 4.4 Commutativity diagram for a coordinate transform

$$\begin{tikzcd}
A \arrow[r, "\gamma"] \arrow[d, "\alpha"'] & B \arrow[d, "\delta"] \\
C \arrow[r, "\beta"'] & D
\end{tikzcd}$$

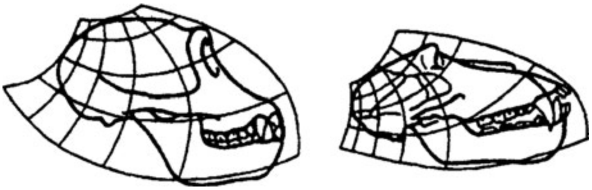


Fig. 4.5 The non-linear transformation of coordinate space shows that a chimpanzee's and baboon's skulls have elements of equivalence (From D'Arcy Thompson, *On Growth and Form*, Reprinted with the permission of Cambridge University Press.)

4.9 Symmetry and Symmetry Operations Allow Molecules to Be Placed in Groups

In chemistry and physics the idea that one system or object may be equivalent to another is usually discussed in terms of *symmetry*. Symmetry is the concept that the properties of systems are unchanged by one or a series of symmetry "operations." In other words, if a molecule or atom can be rotated, translated, reflected, or inverted and the resulting object is indistinguishable from the original, it has a specific symmetry. A square has 9-fold symmetry compared to an 8-fold symmetry for a circle which is invariant through an infinite number of degrees of rotation. Interestingly, the dominant role symmetry plays in modern physical thought has a strongly Aristotelian flavor even though it is widely recognized today that modern science could not evolve until the strangle hold of Aristotelian thought was broken.

Crystals and molecules can be described in terms of the symmetry operations that can be performed upon them. Consideration of the intrinsic symmetries of a molecular system has application in the treatment molecular orbital theory and ligand field theory. The practical result is a greater understanding of bio-organic and organometallic chemistry. Furthermore symmetry has important considerations in the treatment of molecular vibrations and is applied in spectroscopy and in kinetics. These symmetry considerations are applied to finite or discrete objects. An object with a repeating unit has an infinite symmetry because of the translation in space of the unit. This type of symmetry is applicable to lattice structure such as that found in crystals. Since light will interact in a particular way with the repeating units in a lattice, symmetry provides a formal method for exploring the structure of a crystal. This is the theoretical foundation for diffraction analysis in x-ray crystallographic studies.

The formal mathematical study of symmetry is the study of *group theory*. If a collection of elements can be interrelated by a certain set of formal rules we can call those elements a *group*. The groups of interest to the biophysical chemist are those composed of the sets of symmetry operations carried out on crystals and molecules. The study of symmetry is therefore a subset of the overall study of groups. Obviously some molecules have a high degree of symmetry, while others have little or no symmetry. In order to benefit from the mathematical formalisms of group theory we need to have rigid criteria for determination of symmetry. To establish these criteria, a molecule is considered to have certain *symmetry elements*. These symmetry elements are geometrical entities such as a point, line, or plane with respect to which a *symmetry operation* may be carried out. A symmetry operation is a spatial transform that has the following character: After the movement (transform) is carried through, every point of the transformed body is coincident with an equivalent point of the original body. The symmetry operation generates an equivalent configuration that while not necessarily identical to the original is indistinguishable from it. The complete set of symmetry operations that an object or system can be put through constitutes a group. The system of notation used when discussing molecules is given in Fig. 4.6. When discussing symmetry in crystals, the

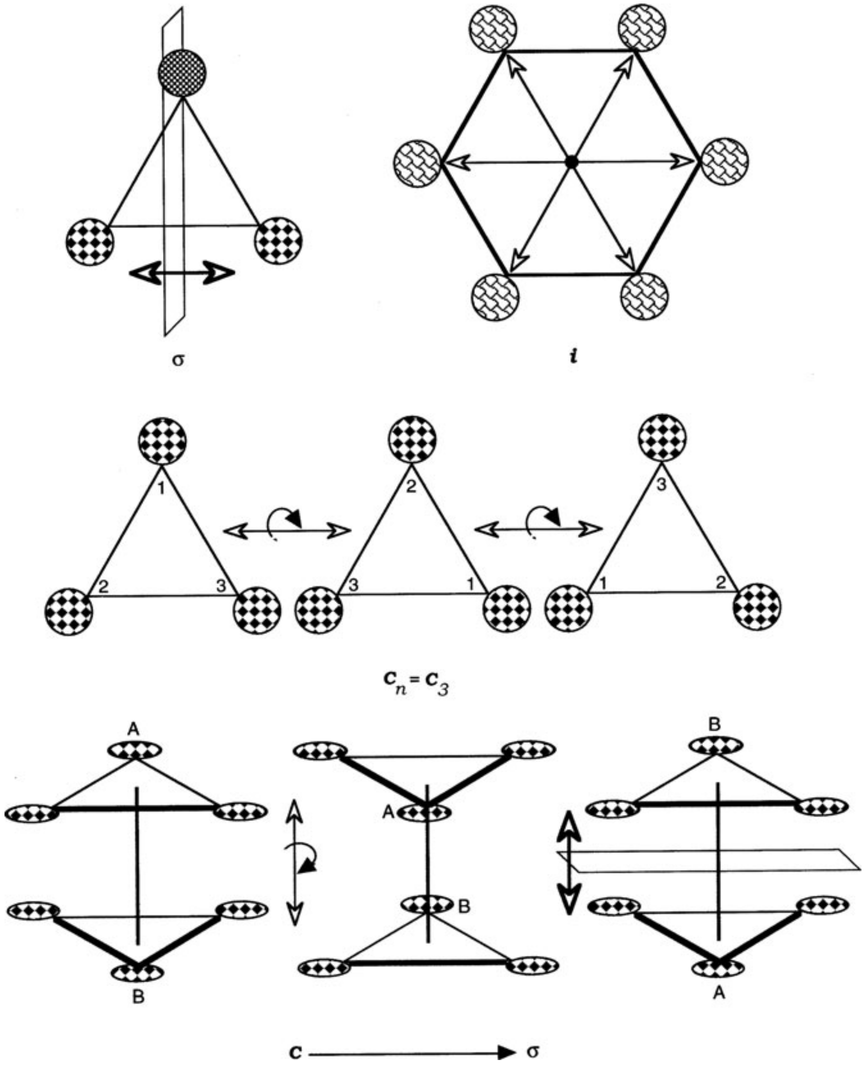


Fig. 4.6 Molecular symmetry elements (and symbols) along with the symmetry operations (and symbols). Not shown is the identity operation called *E*. All molecules if left unchanged have a symmetry element *E*. *Top left* is the *plane* (noted as σ); the associated operation is *reflection in the plane* (also called σ). Successive operations of σ are written σ^n. If *n* is even, $\sigma^n = E$, if *n* is odd, $\sigma^n = \sigma$. *Top right* is a *center of symmetry or inversion* (noted as *i*); the associated operation is *inversion of all atoms through the center* (also called *i*). The effect of carrying out inversion *n* times is expressed as i^n. When *n* is even, $i^n = E$ and when *n* is odd, $i^n = i$. The middle panel shows the element and operation called the proper axis of rotation (*C*). The operation involves *one or more rotations around the axis.* The number of times the smallest rotation operation can be applied to an object to return it to an identical (not just equivalent) configuration is written (C_n). For the triangle illustrated $C_n = C_3$; $C_n^n = E$. Finally the last panel illustrates the rotations that take place around an *improper axis* (*S*). Sequential operations around this axis: rotation followed by reflection in a plane perpendicular to the axis of rotation is improper rotations. The operation of improper rotation by ${}^{2\pi}/_{n}$ is written as S_n

translation of the unit cell in three-dimensional space must be considered in addition to the molecular symmetry factors. We will not discuss these here but details can be found in the references listed at the end of the chapter.

4.10 The Goodness of the Model Depends on Where You Look with Bifurcation Leading to New Discovery

The description of a natural system is inevitably simplified in terms of the observables and linkages between those variables. We usually restrict our efforts to characterize the natural system *N* to a subsystem of *N*. What is needed is some measure of how similar the subsystem is with respect to both *N* and other subsystems in the state space of *N*. We are seeking a measure of "goodness" of our model that will be practically reflected primarily by its dependability as a prediction tool. It is naive to expect the model to fully and faithfully represent the entire state space of N. A certain amount of surprise (unexpected results) and error (incorrectly predicted outcomes) are associated with all modeling processes. These are not errors in measurement nor arithmetic that lead to a surprising or inaccurate result. At issue are systemic errors such as the ultraviolet catastrophe found in the Rayleigh–Jeans law or the surprising result that water expands as it freezes (at least under the conditions of one atmosphere of pressure). These errors and surprises are often described as being at the "fringe of our knowledge" or "on the cutting edge." This choice of language implicitly recognizes that the formal description of errors and surprises is related to points in our state space where a predictive function *bifurcates* (see earlier). A bifurcation exists where an incremental change in the observables that characterize the input and parameters leads to a result that suddenly fits a new and different function. Such a change in function is unexpected. It leads to surprise and a sense of error in regard to the model's goodness. However, the bifurcations are part of the description of *N*, and our surprise is a result of the initial view and constraints we placed on our initial description of the subsystem of *N*.

A simple illustration of a bifurcating function is drawn in Fig. 4.7. When the observable space includes the bifurcation point, surprise is the result. The surprise implies error in the formulation of the description of the original abstract state space. The error represents the limits of the frame in which the system *N* is modeled. A gap exists between reality and what is described in the model. This modeling gap occurs because the model is closed to or does not recognize a set of observables to which the system *N* is open. The gap is characterized by the set of incongruent observables. This "gap set" of incongruent observables is important in the advancement of scientific knowledge. Knowledge is advanced in the scientific method by identifying and elucidating these incongruent observables. The story of how the quantum theory developed (Chapter 8) is an excellent example of the study of a gap set of observables.

Most of the functions that we assign to describe linkages in physical systems have the mathematical property of being smoothly differentiable at least within the regions that *we* define as of interest! Such functions are characterized by a capacity

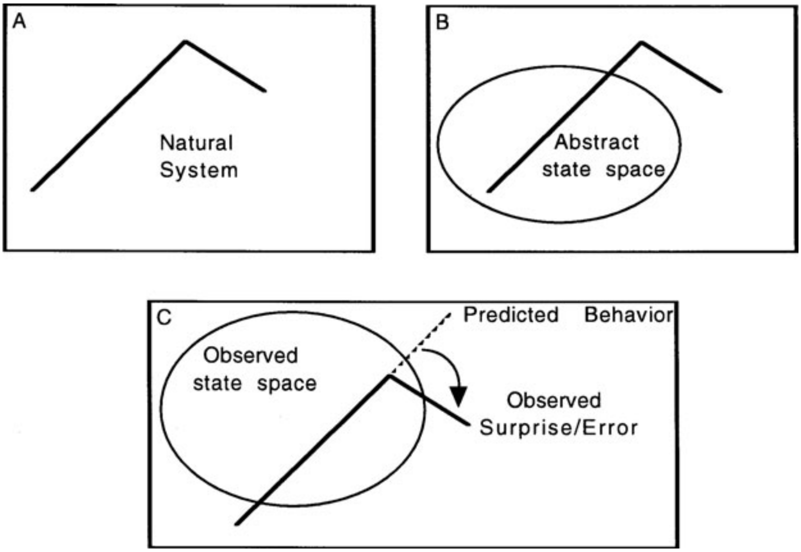


Fig. 4.7 (**a**) A bifurcating function *N*, a natural system. (**b**) The abstract state space used to construct a model of *N*. (**c**) When the observed state space now includes the bifurcation point of *N* a surprise or error occurs because the observed behavior contradicts the predicted behavior

for smooth transformation as we have previously discussed. It is rare for these functions to have bifurcations in which there is a qualitative change in the behavior of the function. The qualitative changes that occur at the bifurcation points lead to non-commutativity and to non-equivalent descriptions. Thus bifurcations represent points in which a new class of system is to be defined. The study of bifurcations is important throughout science but has a growing presence in several fairly distinct areas to which today's student has exposure. We will only briefly mention these here and refer the interested reader the reference list at the end of the chapter for further exploration of these topics.

4.11 Bifurcations in State Space Characterize Complex Systems

A complex system is one in which, despite careful and accurate description of some (or many) subsystems, the overall system behaves in a manner that does not seem to be predicted or accounted for on the basis of those subsystems. In other words, there is no straightforward and simple way to describe what the whole system is doing. This relatively compact intuitive notion of a complexity turns out to be quite difficult to capture formally. If we assume that the complexity of a system is an intrinsic property of the system and is therefore independent of the observer interacting with the system through an abstracted set of observables, the formal description of the system is virtually impossible. What if complexity is not intrinsic to the system but rather a result of the observed:observer relationship?

Let us restate this notion of complexity by recognizing that a complex system comes equipped with results outside of our expectations; there are surprises and errors in the counterintuitive behaviors of the whole system. Such “unpredictable” behaviors of a system are caused by bifurcations in the model that go unrecognized due to closure of observable sets while the system itself remains open to these gap interactions. Error and surprise are indistinguishable from a bifurcation in state space; hence it seems likely that much of a system’s complex behavior results from observer interaction with the system and is not an independent property of the system at all. In this relativistic view, complexity depends on the interaction of one system with another and is not an intrinsic property of the system being observed. The interaction between the two systems, the observing system and the observed system, is mutual and so, in this view, it is not just the independent description of the natural system that matters but the interaction itself is part of the description. In a certain limited set of observations, the interactions between the observer and natural system may be so weak that they may be neglected. In this special case the natural system’s complexity will appear to be an independent and intrinsic property. Such cases are probably limited to physical and engineering systems where classical limits of size and velocity are respected. However, as is confirmed by the collapse of the philosophical basis for classical physics at the end of the nineteenth century, the independence from complexity in even physical systems is an illusion generated by observables through which the natural system is abstracted and described.

A very common example of this type of complexity and surprise in a system is the phase diagram, common in thermodynamics. For example, the triple point of water is a bifurcation point at which any infinitesimal movement away from the triple point leads to a completely different set of properties associated with the different phases, ice, liquid, and steam (see Fig. 13.1). As we will see in Chapter 14, many of the very unique properties of water can be deduced because of the “surprises” that govern the state space near the triple point. Similar examples of complexity, surprise, and error and the window they provide to important systems abound in the biological world.

Another example of surprise in state space that is of great practical consequence is the difficulty encountered by the biotechnology industry in making a biologically active gene product. The surprise that a properly sequenced primary polypeptide may not proceed to the correct structure and function is the result of our expecting that primary sequence dictates secondary and tertiary structure. As we will see in Chapter 19, the observables as well as the equation of state on which we base these expectations are not as well developed and known as we would like. Furthermore there are post-translational events such as glycosylation and myristolization, which clearly play a role in the structure–function equation of state but which are not included in the manufacture of a biotechnologically generated molecule. From a systems standpoint the failure to make the product with the anticipated activity is not surprising. It is predictable because the natural system has many important and probable bifurcations that serve to delineate our ignorance that must be dependably characterized in order for our knowledge actually to advance.

4.12 Catastrophes and Chaos Are Examples of Formal Mathematical Systems That May Capture Important Behaviors of Natural Systems

Most interesting physical problems involve change. The modeling of change in a system is called *dynamical modeling*. Here we consider the movement of a subsystem with respect to time in a space of all possible states. Most models of natural systems use smooth and continuous mathematical functions to describe the systems. Elementary functions such as cos x, sin x and e^x are continuously differentiable. But much of the change in the natural world is discontinuous. The assumption that changes occur in a continuous manner such that x plus Δx is essentially the same as x (which is the assumption of classical physics) does not hold well. In the realm of quantum mechanics the assumptions of continuity are discarded, but even in the functional classical realm; beams buckle, atmospheric turbulence leads to tornadoes, and cell differentiation commonly occurs. All are singular or discontinuous events. The ability to describe systems replete with discontinuity through the mathematics of smoothly varying functions is the goal of bifurcation theory. Catastrophe and singularity theory are related fields of investigation. We touch briefly on catastrophe behavior because of its initial application to biological systems. Catastrophe behavior is usually seen in both formal and natural systems when a smoothly varying function is suddenly found to become discontinuous.

The central question in catastrophe theory is what a particular observable will look like in the region of a critical point. This is where the question of one function being similar to another function becomes important. If one function can be transformed into another by a smooth coordinate change and will therefore commute, then the functions look alike and are essentially equivalent. When an observable takes on a value in parameter space such that a small change does not give a smooth coordinate change and the function does not commute, then these values represent a bifurcation point. The set of these values is called the *catastrophe set*. When certain functions take on values in the catastrophe set, the behavior of the function (or observable) will be partitioned into distinctly different disjoint regions of state space and these will have a substantially different phenotype when compared to the other regions in the state space. As can be imagined, this is the description of an edge or cusp in state space. If we imagine walking along such a cusp (i.e., if we allow the parameters of the observable (observer) to take on a value in the catastrophe set or manifold), then we can easily see that taking a value of a parameter in normal space, either stepping left or right, will lead rapidly to two distinct and non-commutable states. The structure of this elementary catastrophe called the *cusp* is shown in Fig. 4.8.

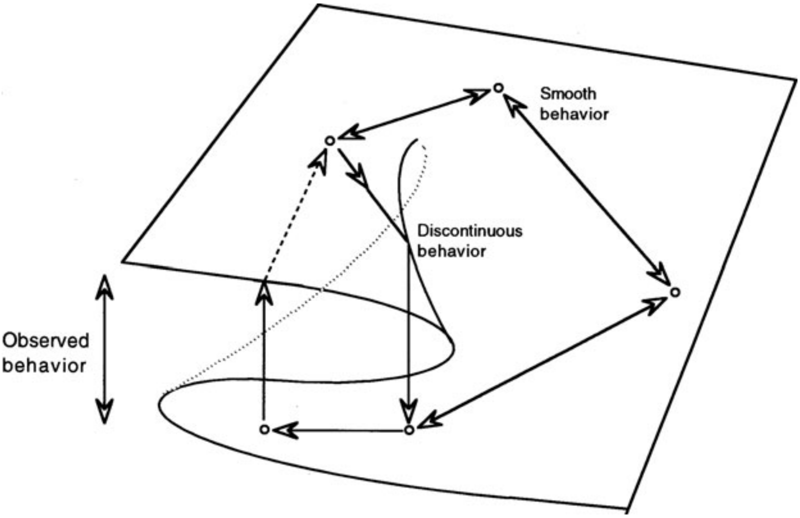


Fig. 4.8 The structure of the elementary catastrophe cusp

Other structures are associated with catastrophe sets found with higher order polynomials and gives rise to such descriptive names as “butterfly” and “swallowtail.” On a historical note, the mathematician René Thom raised the intriguing contention that catastrophe surfaces could be found in biological systems especially during development. For example, Fig. 4.9 shows in parallel the formation of a pocket in the developing sea urchin and in the elliptic umbilic catastrophe. It has been found that many complicated physical processes can be described in terms of catastrophe sets. This makes the modeling of bifurcating systems of great interest both as a tool and as an epistemological lens into the study of these systems.

Chaotic behavior is an extension of the study of dynamical systems and is important in the control of some interesting systems. First we must define several terms used extensively in dynamical modeling. In dynamical modeling a geometric approach is taken to draw a *phase portrait* of a system. To construct a phase portrait, a geometric model of all the possible states for the model is drawn; this is called the *state space*. Into this state space are drawn dynamically determined structures which are comprised of *basins* or wells separated by higher walls called *separatrices*. The nuclear structure of each basin is the *attractor.* Of all the possible abstract states described by a model, it is the attractors that are the observed states. From a dynamical viewpoint, if an observational system (an experimental device) is placed in the state space it will move with a certain trajectory toward a dynamic equilibrium that is a stable limit set. The experimental device can be dropped anywhere in the state space and many trajectories can be described all moving in some relation to the stable limit set. Some trajectories may not move toward the limit set but may pass it or even move away from it. Only the stable limit sets that receive most of the trajectories will be observed. The stable limit sets are the attractors of the system. If all

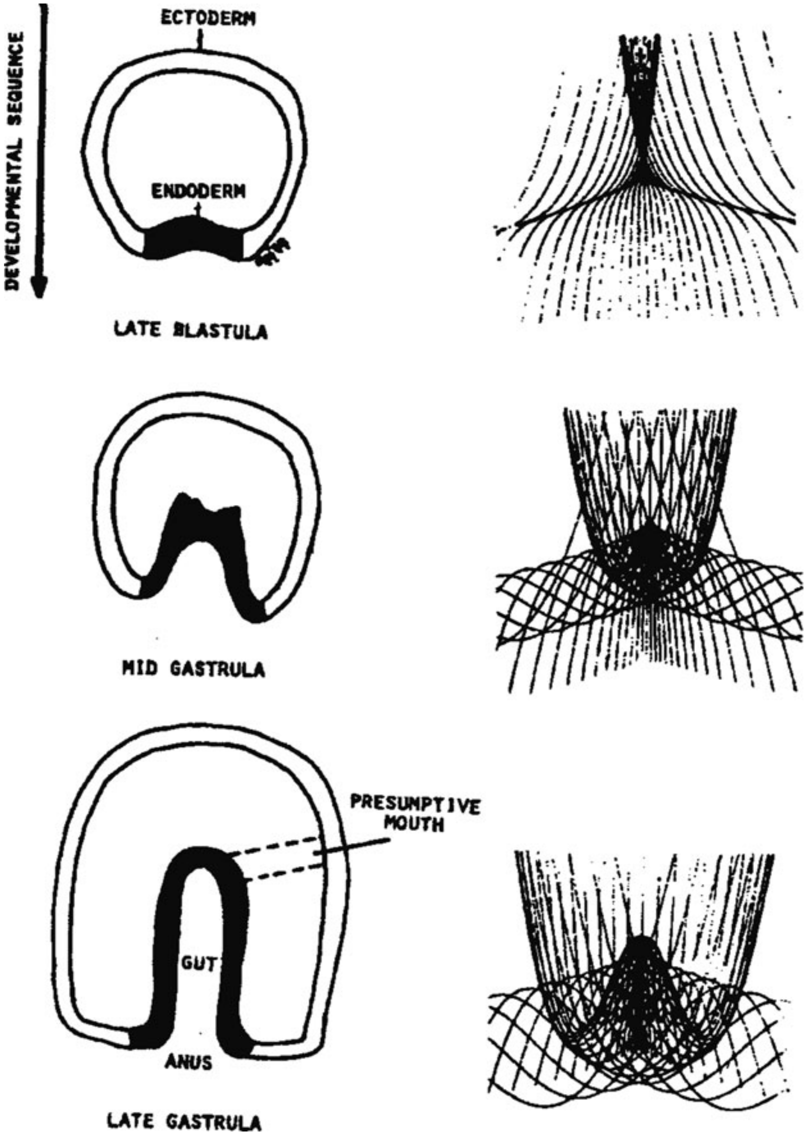


Fig. 4.9 (**a**) The formation of a pocket in the developing sea urchin and (**b**) the similar dynamic pattern of the elliptic umbilic catastrophe surface (From *Catastrophe Theory* by Alexander Woodcock and Monte Davis. Copyright © 1978 by Alexander Woodcock and Monte Davis. Used by permission of Dutton Signet, a division of Penguin Books USA Inc.)

the trajectories that can be observed arrive at a single point, the attractor is called a *static attractor*. Some trajectories may tend to settle into a cyclical trajectory which appears to have a stable "orbit" around a critical point. This stable trajectory is a limit cycle and is also an attractor called a *periodic attractor*.

In a deterministic model we would expect that, given the description of the state space and its cellular elements comprised of attractors, basins and separatrices, we could drop our experimenter into the state space and with a single look know where the experimental package is and thus predict the future along the trajectory with complete certainty. It turns out that *at least experimentally* this is not always the case. Certain attractors have been experimentally found to have a much more complex local structure than a single point or a stable orbit. The trajectories that make up these attractors describe a complicated space that acts as if it has many local bifurcations (i.e., many surprises). This is a *chaotic attractor*. These three attractors are illustrated in Fig. 4.10. A trajectory that ends in a chaotic attractor will appear to have a very unpredictable and turbulent "dynamic equilibrium." This experimentally unpredictable observed behavior is the origin of the term "chaotic." It is important to strongly emphasize that the connection between chaotic attractors of mathematical theory and the attractors seen experimentally is an area of active research. A trajectory that orbits in a chaotic attractor will be experimentally unpredictable for several reasons, especially the impossibility of knowing the exact position of the trajectory at any given moment (indeterminacy). The results are that a small difference in present position will lead to an enormous difference in position in the future and that the trajectory can occupy (and at some point in time will occupy) every region in the complicated local space of the chaotic attractor.

It is apparent why chaotic systems, which seem to be apt descriptions of the observed behavior of turbulent systems and can give rise to an enormous variety

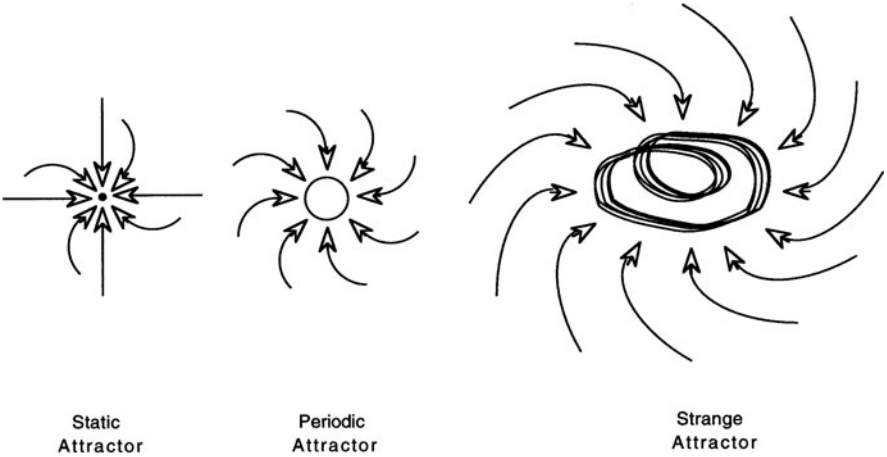


Fig. 4.10 Attractors. The figure on the left is a static attractor, and the one in the middle is a periodic attractor. Both are stable systems. These are both classical systems in which the future behavior of all approaching vectors can be known with 100% certainty following a single measurement. The attractor on the right is a chaotic attractor, sometimes called a strange attractor. It is impossible to predict the future behavior of vector, in state space with absolute precision based on a one-time measurement

of states that are closely associated are popular for study of biological and biophysical phenomena. (For example, consider the structure of a protein and all of its metastable states as a chaotic attractor).

"The simpler and more fundamental our assumptions become, the more intricate is our mathematical tool of reasoning; the way from theory to observation becomes longer, more subtle and more complicated." Albert Einstein and Leopold Infeld in *The Evolution of Physics* (New York: Simon and Schuster, 1950).

Further Reading

General Modeling

Casti J.L. (1992) *Reality Rules: I; Picturing the World in Mathematics – The Fundamentals*. Wiley, New York. (This volume is a delightful introduction to the intellectual approach to modeling. This chapter has used Casti's approach in a simplified fashion and his book is the logical next step to expand your horizons.)

Casti J.L. (1992) *Reality Rules: II; Picturing the World in Mathematics – The Frontier*. Wiley, New York. (Going further than the first volume, this exposition is just as much fun and intellectually satisfying as the first volume.)

Biological Modeling

Murray J.D. (1991) *Mathematical Biology*, 2nd edition. Springer-Verlag, New York. (This textbook devotes a detailed chapter to a selection of 20 fundamental processes of biological interest. Each chapter describes and explores these models. Included are models important to our study such as waves of diffusion, population growth, reaction mechanisms, enzyme kinetics, biological oscillators etc. Detailed and from the mathematician's point of view.)

Symmetry

Atkins P.W. (1983) *Group Theory, Chapter 19* ***in*** *Molecular Quantum Mechanics*. Oxford University Press, Oxford. (A good introduction to group theory and symmetry. PW Atkins also treats the subject nicely in his textbook, *Physical Chemistry*.)

Cotton F.A. (1990) *Chemical Applications of Group Theory*, 3rd edition. Wiley, New York. (For the serious student. This is a book worth studying but can not be picked up casually.)

Feynman R.P., Leighton R.B., and Sands M. (1963) *Symmetry in Physical Laws, Lecture #52* ***in*** *The Feynman Lectures on Physics,* Volume 1. Addison-Wesley, Reading, MA. (Clear, concise, illuminating. the first stop for those going further.)

Dynamical Systems

Abraham R. and Shaw C.D. (1984a) *Dynamics, The Geometry of Behavior; Part One: Periodic Behavior*. Aerial Press, Santa Cruz, CA.

Abraham R. and Shaw C.D. (1984b) *Dynamics, The Geometry of Behavior; Part Two: Chaotic Behavior.* Aerial Press, Santa Cruz, CA.

(These two volumes (there are four in the set) treat dynamical modeling in graphical and remark ably jargon/symbol free fashion. These books are designed to promote "mathematical literacy" with the belief that mathematical ideas are fundamental to any meaningful the philosophical dis course in a culture. They are an excellent starting point for the biologist afraid of "advanced" topics.)

Beltrami E. (1987) *Mathematics for Dynamic Modeling*. Academic, Boston. (A more typical mathematical treatment of the above subjects. Oriented to specific modeling problems.)

When You Want More

Ditto W.L. and Pecora L.M. (1993) Mastering Chaos, *Sci. Am.*, **269, 2**:78–84. (Non-technical general introduction to chaos theory.)

Ekeland I. (1988) *Mathematics and the Unexpected*. University of Chicago Press, Chicago. (This volume discusses Kepler, chaos and catastrophe and is a logical next step.)

Thom R. (1975) *Structural Stability and Morphogenesis*. W.A. Benjamin, Reading, MA. (This volume is by the originator of the catastrophe theory.)

Thompson, D. W. (1942) *On Growth and Form*. Cambridge University Press, Cambridge. (This is his classic, revolutionary work on evolution and coordinate transformations.)

Weiner N. (1961) *Cybernetics or Control and Communication in the Animal and the Machine*, 2nd edition. MIT Press and Wiley, New York. (This is a classic and should be required reading for any scientist. No longer out of print and now widely available. Find the professor with it on his/her shelf and talk with her/him.)

Woodcock T. and Davis M. (1978) *Catastrophe Theory*. Dutton, New York. (A good introduction to catastrophe theory for the non-mathematician.)

Problem Sets

1. Compare the pre-Copernican model of geo-centricism and the subsequent model of heliocentricism in terms of the coupling assumptions between observer and the observed.
2. List several observables that will provide information about the metabolic state of a liver cell.
3. List three observables characterizing the state of a muscle cell in terms of its metabolic activity.
4. Describe the linkages between the observables chosen in questions 2 and 3. Arrange the observables and linkages into an equation of state that reflects the metabolic state of the system.
5. In modern MRI (magnetic resonance imaging) and SPECT (single photon emission computerized tomography) scans the instruments measure the amount of

blood flowing to a specific area of the brain in order to assess the "intellectual" use of that part of the brain.

What are the abstractions of this process that allow conclusions to be drawn? Are there likely to be any bifurcations or surprises in the linkages in your proposed state space?

6. The PET (positron emission tomography) scanner uses a tagged radioisotope to measure glucose delivery. Is this a better system for observing "intellectual" function than those described in question 4?
7. Superoxide dismutase is family of enzymes that protects cells against free radical damage by $O_2^{\bullet-}$. They catalyze the following reaction:

$$O_2^{\bullet-} + O_2^{\bullet-} + 2H^+ \rightarrow O_2 + H_2O_2$$

Superoxide dismutase has been isolated from both prokaryotic and eukaryotic organisms. Under what conditions can SODs be considered equivalent? What observables would you choose to argue your case?

8. Describe the symmetry elements and symmetry operations for the following:

a. circle
b. pentagon
c. Tyrosine side chain
d. Hexagon
e. Hexagonal array

9. A prominent biochemist has been quoted as arguing in curriculum discussions that: "a *modern biochemist does not need to know any biophysical chemistry in order to be successful"*. Without quibbling over the term successful, explain why such a statement may be regarded as true or false. Proceed with your analysis in terms of systems theory and explain your reasoning with precise explication of the state space you are discussing, its observables, linkages, errors and surprises.

(Hint: whatever viewpoint you choose to argue, the argument will be more easily made if you make a specific biological reference and relate your formal argument to a real system.)

Chapter 5
Probability and Statistics

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_5, © Springer Science+Business Media, LLC 2010

5.1 An Overview of Probability and Statistics

Few subjects in mathematics and science are more widely used, feared, and scorned than probability and its derivative discipline statistics. Defining probabilities is a fundamental process in the observation of state space and forms the basis for much of modern scientific theory and experimentation. We will use probability functions throughout this book, and we will review most of the concepts and distributions needed for the consideration of most of these areas of study in this section.

Probability problems are either *discrete* or *continuous*. If an experimental situation has a countably limited number of possible outcomes, then the problem is discrete. It is important to recognize that the countable number may be infinite, but as long as the possible outcomes are *countably* infinite it will remain a discrete problem. Alternatively, when a problem has an uncountably infinite set of possible outcomes then the problem is continuous. An example of a discrete problem in probability is to determine how many times a die could be rolled before getting a "six." Since there are a discrete and countable number of outcomes, this problem is treated by discrete techniques (even though it is possible to roll the dice an infinite number of times and never roll a "six"). Here the random variable can only take on a limited number of variables. When the random variable can take on any value in an interval, however, the problem becomes one of continuous probability. An example is the point in time at which two ants crawling around a room will meet one another at the center of the room between 3 and 4 PM. There are an infinite number of times that the first ant could be at the center of the room and these times are not countable since they cannot all be listed and labeled $\{a, b, c, \ldots\}$. (One can always select another time between any other two chosen times, thus making the possibilities both infinite and uncountable.) The second ant can also arrive at the center of the room (an infinite number of possible times). This problem is therefore continuous. The mathematics used to solve discrete problems is simple algebra, but continuous problems will require the use of integral calculus.

5.2 Discrete Probability Counts the Number of Ways Things Can Happen

Probability is a concept linked to an experiment or trial that can be repeated many times and for which the frequency or occurrence of a particular outcome is of interest to us. The probability of outcome, A, is the fraction of times that A occurs in a set of trials. An experiment possesses a sample space that contains the set of outcomes such that exactly one outcome occurs when the experiment is performed. The outcomes are points in the sample space and the sample space may be defined in a wide variety of ways. Recall our previous discussion of state spaces and the ideas of choosing appropriate observables. We took care with abstractions so that the required observables necessary to describe the links between formal and real systems were included rather than being discarded. There is often more than one way to view an experiment. James Clerk Maxwell said, “the success of any physical investigation depends on the judicious selection of what is to be observed as of primary importance.” As an example, consider a shelf of 25 books and the random drawing of a single book from the shelf. Some of the sample spaces describing the set of outcomes are as follows:

1) The space consisting of 25 points, each representing one of the books.
2) The space consisting of books with green covers.
3) The space consisting of books with no mathematical formalisms in the text.
4) The space consisting of books on biophysical chemistry (hopefully this set is the same as set #1 in your library!).
5) Etc.

An *event* is defined as any outcome or collection of outcomes in a given sample space. An event may include the null set as well as the set of all outcomes.

Each event is assigned a likelihood of occurring if a trial is performed. Consider an experimental situation in which we will measure the likelihood of a dart thrown from 5 ft hitting a small circular dartboard. If the probability of the thrown dart striking the board is 0.18, we would expect the dart to strike the board 18% of the times it is thrown. In the first 100 trials, however, we expect the dart to strike the board approximately 18 times. If many more trials are attempted, the chance of a precise 18% strike rate emerges because of *statistical convergence*. If in 10 million trials the strike rate is 18% it is quite accurate to expect a strike rate for the second 10 million trials to be 18% or a rate extremely close to this. A well-defined probability of an outcome, A, is the fraction of trials in which A occurs when the number of trials is very large. A probability is a number that represents this fraction and can be assigned when an event satisfies these rules:

1) An outcome given is a non-negative probability such that the sum of all the n probabilities equals one. (Thus the total of all possible outcomes equals the certainty that something *will* happen.)
2) If A is an event and $P(A)$ represents the probability of A, then $P(A)$ equals the sum of the probabilities of the outcomes in the event A. With the definition of an event as given above, $P(A)$ may represent a single outcome or a group of outcomes depending on how the event is defined.

These rules allow us to define a *probability space* as a sample space together with the assigned probabilities of the outcomes. Figure 5.1 shows an example of a probability space A in which an event E is defined containing five outcomes: a, b, c, d, and e. The value of $P(E)$ is given by the following:

$$P(E) = 0.1 + 0.2 + 0.15 + 0.1 + 0.2 = 0.75 \tag{5.1}$$

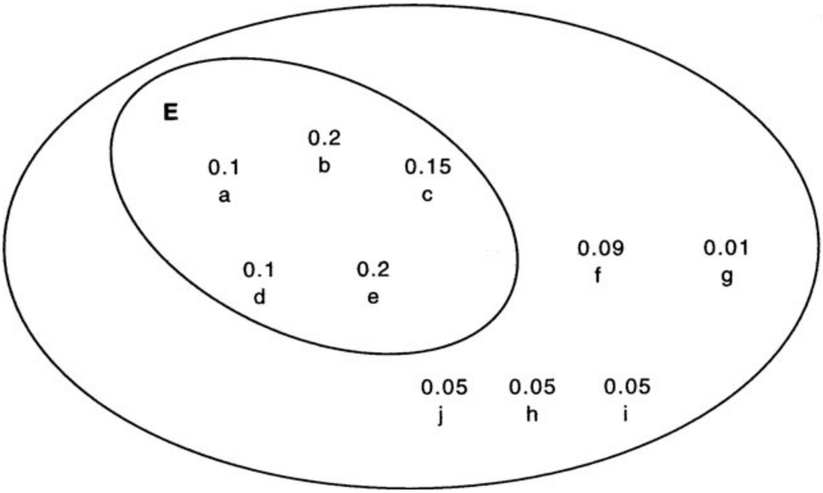


Fig. 5.1 A probability space. The *large circle* is the probability space that contains 10 points each with an assigned probability. The event, *E*, is represented by the *smaller circle* and contains five points whose sum is 0.75

An important special case exists in which all outcomes have an equal probability (i.e., an experiment is *fair*) and in which the outcomes are picked at random. Then, if a sample space has *n* points:

$$P(\text{each outcome}) = \frac{1}{n} \text{and}\, P(E) = \frac{\text{favorable outcomes}}{n} \tag{5.2}$$

We can (informally) define some terms and concepts:

1) A *sure event* is one containing all points in the sample space, its probability $P(E) = 1$.
2) An *impossible event* contains no points in the sample space, its $P(E) = 0$.

Converse statements of 1 and 2 are not necessarily true.

A *complementary* event may be defined in terms of E: if E is an event, then $\overline{E}$ is the set of outcomes not in E, thus

$$P(E) = 1 - P\left(\overline{E}\right) \tag{5.3}$$

5.3 Specific Techniques Are Needed for Discrete Counting

The most common probability space we will consider will be the space in which equal probability outcomes (a fair experiment) are randomly chosen. It is not always simple to count each outcome and so a variety of counting techniques are valuable including the principles of multiplication, permutation, combination, distinguishable and indistinguishable outcomes and symmetries.

5.3.1 Multiplication Counts Possible Outcomes of Successive Events

When considering the total number of possible outcomes for an event that takes place in successive stages, each stage is considered to have a certain number of slots or boxes. The number of slots for each stage is multiplied together to give the total number of possible outcomes.

Example 1: The total number of tetrapeptide sequences is

$$20 \times 20 \times 20 \times 20 = 20^4 \tag{5.4}$$

since each slot in the sequence can be filled by one of 20 possible amino acids.

In this example each amino acid can be used over and over again because the pool of the amino acids is continuously replaced, this is called *sampling with replacement*.

The case in which *sampling without replacement* occurs is also important.

Example 2: The total number of tetrapeptide sequences that can be formed from a pool of 20 amino acids is

$$20 \times 19 \times 18 \times 17 \tag{5.5}$$

The first spot can be filled 20 ways but only 19 possibilities exist for the second slot since only 19 amino acids remain, etc.

Problems of sampling without replacement are so common that the notation for the factorial can be introduced. The product $n!$ (n factorial) is defined as:

$$n! = n(n-1)(n-2)\cdots 1 \tag{5.6}$$

$$0! \text{ is defined to be} 1 \tag{5.7}$$

Using this notation, example 2 above can be written as

$$\frac{20!}{16!} = 20 \times 19 \times 18 \times 17 = 116,280 \tag{5.8}$$

5.3.2 Permutations Are Counts of Lineups

The number of ways n objects can be ordered or lined up one at a time is $n!$. In general, this is simply the problem of sampling without replacement. For example, consider a test tube containing one each of the five nucleotides contained in nucleic acids. What are the possible permutations of these five nucleotides? Each nucleotide can be placed into one of five slots but cannot be reused once it occupies a slot so the total number of permutations is $5 \times 4 \times 3 \times 2 \times 1$ or $5!$.

If the problem involves the number of permutations of n distinct objects taken r at a time, the expression is

$$nPr = \frac{n!}{(n-r)!} \tag{5.9}$$

This formula gives the number of permutations when a population of size n is sampled in groups of size r, e.g., we read permutations of 20 take 4, e.g.,

$$nPr = \frac{20!}{(20-4)!} = \frac{20!}{16!} = 116,280 \tag{5.10}$$

5.3.3 Combinations Are Counts of Committees

The key difference between a permutation and a combination is that the order of the elements in a combination does not matter, whereas it did matter for the permutation. In a combination, there is no first place, second place, and so on, so the combinations *abcd* and *adcb* are identical and are not counted twice. If we wish to find the combinations of size 3 out of 20 objects and were to write these out explicitly, the relationship between permutations and combinations would be found to be that each combination will give rise to 3! permutations of that group of elements such that *combinations* $\times\ 3! =$ *permutations*. Thus,

$$\#\ \text{combinations} = \frac{\text{permutations}}{3!} = \frac{20!/17!}{3!} = \frac{20! \times 19! \times 18!}{3!} \tag{5.11}$$

The notation for this is $\binom{20}{3}$ which stands for combinations of size 3 from a population of size 20. The notation is read as *20 on 3* or *20 choose 3*. More generally, this is a binomial coefficient and $\binom{n}{r}$ is the general notation for the number of combinations of n elements taken r at a time. This may be evaluated as

$$\binom{n}{r} = \frac{n!}{r!(n-r)!} \tag{5.12}$$

5.3.4 Counting Indistinguishable Versus Distinguishable Entities Require Different Techniques

Problem: We wish to synthesize a protein and have a mixture that contains 50 alanines, 60 glutamines, and 70 lysines. Pick 20 sequential amino acids without replacement and find the probability of getting 10 lysines, 4, alanines, and 6 glutamines.

Solution: The total number of possible outcomes is $\begin{pmatrix}180\\20\end{pmatrix}$. For the favorable outcome sought, pick 10 lysines out of 70, pick 4 alanines out of 50, and 6 glutamines out of 60, thus,

$$P(10\text{ lys}, 4\text{ala}, 6\text{gln}) = \frac{\begin{pmatrix}4\\50\end{pmatrix}\begin{pmatrix}6\\60\end{pmatrix}\begin{pmatrix}10\\70\end{pmatrix}}{\begin{pmatrix}180\\20\end{pmatrix}} \tag{5.13}$$

This treatment assumes that the amino acids in the box can be distinguished from one another and so each can be named as lys_1, lys_2, . . ., lys_{70}, etc. The probability of getting an amino acid does not change whether or not it is named.

There are cases in which the elements are not able to be distinguished and a different counting method must be used. The counting of indistinguishable cases will be introduced when we discuss statistical thermodynamics in Chapter 11.

5.4 Counting Conditional and Independent Events That Occur in Multistage Experiments Require Special Considerations

Many situations have more than a single stage and such experiments may be regarded as one of two types:

a) *Conditional* events, where the outcome of the second stage depends on the result of first stage.
b) *Independent* events in which the second-stage outcome has no dependence on the first stage.

Conditional probabilities occur in problems in which we calculate the probability of B occurring given that A has already occurred, $P(B|A)$. For example, in making a bipeptide from a pool of 20 amino acids (without replacement):

$$P(\text{aliphatic a.a. in position2}|\text{aliphatic a.a. in position 1}) = \frac{4\text{aliphatics left}}{19\text{a.a.left}} \tag{5.14}$$

$$P(\text{aliphatic a.a. in position2}|\text{aromatic a.a. in position 1}) = \frac{5\text{aliphatics left}}{19\text{a.a.left}} \tag{5.15}$$

There are always only a maximum of five aliphatic amino acids available [leucine, isoleucine, valine, alanine, glycine]. In the first case, once one of these are used, only four choices remain. In the second case, the placement of tyrosine, phenylalanine, or tryptophan in the first position leaves all five of the aliphatic choices available for the second position.

In general, conditional probabilities are defined as

$$P(B|A) = \frac{P(A \text{ and } B)}{P(A)} = \frac{\text{favorable outcome in A space}}{\text{total probability of the A space}} \tag{5.16}$$

Consider the two probability spaces A and B in Fig. 5.2. $P(A) = 0.8$ and $P(B) = 0.3$. When we determine $P(B|A)$, A now becomes the new universe and just a single outcome for B exists, 0.2. This probability is now more significant because the total probability of A is 0.8 (with respect to the former universe). Thus Eq. (5.14) gives

$$P(B|A) = \frac{0.2}{0.8} \tag{5.17}$$

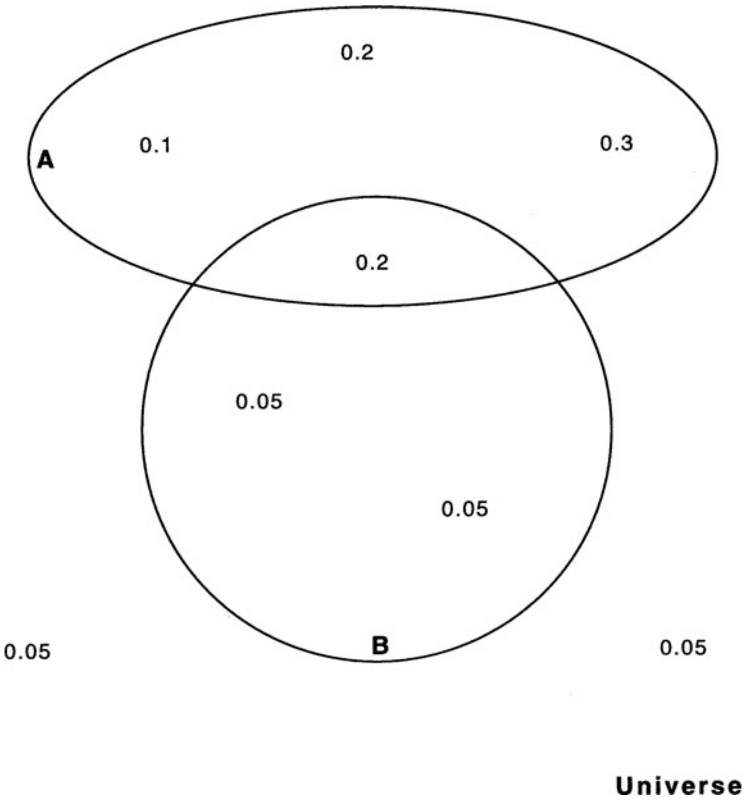


Fig. 5.2 The probability space in a conditional case

Equation (5.14) can be written as an AND rule:

$$P(A \text{ and } B) = P(A)P(B|A) \tag{5.18}$$

which, by symmetry, gives

$$P(A \text{ and } B) = P(B)P(A|B) \tag{5.19}$$

Multiple conditional events can be determined by the chain rule:

$$P(A \,\&\, B \,\&\, C) = P(A)\ P(B|A)\ P(C|AB) \tag{5.20}$$

Events are independent when the occurrence (or lack of occurrence) of one event has no impact on the probability of the second event:

$$\text{iff } P(A \text{ and } B) = P(A)P(B), \text{then } A \text{ and } B \text{ are independent} \tag{5.21}$$

Two other tests of independence arise if Eq. (5.21) is combined with Eq. (5.18) or (5.19):

$$\text{iff } P(B|A) = P(B),\ A \text{ and } B \text{ are independent} \tag{5.22}$$

$$\text{iff } P(A|B) = P(A), A \text{ and } B \text{ are independent} \tag{5.23}$$

5.5 Discrete Distributions Come from Counting Up the Outcomes of Repeated Experiments

5.5.1 The Multinomial Coefficient

Many physical problems are concerned with the case of a discrete experiment repeated over and over when the multiple stages are independent. We will first look at a specific example and then write a general formula that calculates the probability of finding a particular distribution of outcomes. Consider an experiment with four mutually exclusive outcomes, A, B, C, and D. If the experiment is repeated 10 times, the probability of the distribution $P(3A, 2B, 3C, 2D)$ is given by

$$P(3A,\ 2B,\ 3C,\ 2D) = \frac{10!}{3!2!3!2!}\left[P(A)^3][P(B)^2][P(C)^3][P(D)^2\right] \tag{5.24}$$

The sample space for this problem consists of all the strings of length 10 using four outcomes (in this case called A through D), 4^{10}. A priori, the 4^{10} outcomes are not equally likely since *BBBBBBBBBB* has a probability of $[P(B)^{10}]$ and *CCCCCCCCCC* has a probability of $[P(C)^{10}]$. Because of this a probability based on $\dfrac{\text{favorable outcomes}}{\text{total outcomes}}$ can not be used; instead, the probability of each favorable outcome is determined and then these are all added together. A favorable outcome is represented by *AAABBCCCDD* and since each trial is independent, we can write

$$\begin{aligned} P(AAABBCCCDD) =&P(A)P(A)P(A)P(B)P(B)P(C)P(C)P(C)P(D)P(D) \\ =&[P(A)^3][P(B)^2][P(C)^3][P(D)^2] \end{aligned} \tag{5.25}$$

Write out as many expressions for favorable outcomes as you need to be convinced that all of these outcomes will have the same probability as Eq. (5.25).

We have now calculated the probability $P(\text{favorable outcome}) = P(3A, 2B, 3C, 2D)$ which is the second term in Eq. (5.24) and now we need to now how many favorable outcomes there are. In other words, how many ways can $3A$, $2B$, $3C$, $2D$ be arranged? Which is asking, What are the permutations? If each result could be named $(A_1, A_2, A_3, B_1, B_2, C_1, C_2, C_3, D_1, D_2)$ there would be 10! permutations. However, the results are identical and this approach leads to overcounting *A* by 3!, *B* by 2!, *C* by 3!, and *D* by 2!. Therefore the number of permutations is $\frac{10!}{3!2!3!2!}$, which is the first term of Eq. (5.24). This example can be generalized to *N* independent trials, where each trial has *r* possible outcomes:

$$P(N\,\text{into}\,r\,groups) = \frac{N!}{\prod_{j=1}^{r} N_j!} \tag{5.26}$$

This is called the multinomial coefficient since it occurs in the expansion for a multinomial expression, $(x_1 + x_2 + \cdots + x_r)^N$:

$$(x_1 + x_2 + \cdots + x_r)^N = \sum_{N_1=0}^{N} \sum_{N_2=0}^{N} \cdots \sum_{N_r=0}^{N} \frac{N! x_1^{N(1)} \cdots x_1^{N(r)}}{\int_{j=1}^{r} N_j!} \tag{5.27}$$

where the restriction applies that $N_1 + N_2 + \cdots + N_r = N$.

Some independent trial cases in which the multinomial coefficient is useful in finding the probability distribution are as follows:

1) Tossing dice
2) Coin tosses
3) Drawing cards with replacement
4) Drawing from a very large population without replacement (this condition is not actually independent but if the population is large enough, then one drawing has little effect on the next and to a reasonable practical approximation the drawings are independent of one another)

Furthermore, Eq. (5.26) is important for consideration of systems studied in statistical thermodynamics.

5.5.2 The Binomial Distribution Captures the Probability of Success in the Case of Two Possible Outcomes

If an experiment has two possible outcomes, called either success or failure (e.g., light on or off, engine starts or does not, Na^+ gate is open or closed), then repetitions of this independent experiment are called *Bernoulli trials*. We can write the probability for any single trial as P(success) equals p and P(failure) equals $1 - p = q$. Now we repeat the experiment in a series of Bernoulli trials. The probability of success in these trials can be found as a special case of the multinomial distribution:

$$\begin{aligned} P(k \text{ successes in } n \text{ trials}) &= P(k \text{ successes, } n-k \text{ failures}) \\ &= \frac{n!}{k!(n-k)!} p^k q^{n-k} \end{aligned} \tag{5.28}$$

The coefficient should look familiar from our earlier discussion of combinations and can be rewritten as $\binom{n}{k}$ or the binomial coefficient. Equation (5.28) can now be rewritten in terms of this coefficient and is called the *binomial distribution*:

$$P(k \text{ successes in } n \text{ trials}) = \binom{n}{k} p^k q^{n-k} \tag{5.29}$$

Each outcome consisting of k successes and n minus k failures has a probability of $p^k q^{n-k}$. The binomial coefficient determines how many of these outcomes there will be. The number of successes in n Bernoulli trials (P(success) $= p$) is said to have a binomial distribution with the parameters of n and p.

For completeness, we note that like the multinomial coefficient, the binomial coefficient is so named because it is found in the binomial expansion:

$$(x+y)^N = \sum_{N_1=0}^{N} \frac{N!}{N_1!(N-N_1)!} x^{(N-N_1)} y^{(N_1)} \tag{5.30}$$

5.5.3 The Poisson Distribution Requires Fewer Parameters for Calculation Than the Binomial Distribution

The binomial distribution has a wide range of applications but has the disadvantage of being defined by two parameters n and p which must be known. A related formula that in many cases reasonably approximates the binomial distribution is the *Poisson distribution*. The Poisson distribution has the advantage over the binomial distribution of being characterized by a single parameter λ. The Poisson distribution is written as follows for a λ of fixed value and an experiment with outcome, $k = 1, 2, 3$:

$$P\,(\text{outcome is}\,k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{5.31}$$

The Poisson distribution (which is shown to be a legitimate function in Appendix D) can be used to good approximation to replace the binomial distribution in cases where we cannot apply the binomial because we do not know the parameters n and p.

First we will show that the Poisson distribution is a fair approximation to the binomial distribution. An experiment is performed of n Bernoulli trials such that $P(\text{success}) = p$. These parameters may not always be known (see the following list) but the average number of successes expected in n trials may be more accessible. Thus, we write $np = \lambda$. Within certain limitations

1) n is large i.e., (large number of trials)
2) p is small i.e., (successes are rare)
3) the product λ is moderate
4) $k << n$

we can show

$$P(k\,\text{successes in}\,n\,\text{trials}) = \binom{n}{k} p^k (1-p)^{n-k} \approx \frac{e^{-\lambda}\lambda^k}{k!} \tag{5.32}$$

(The term to the left of the approximation sign is the binomial and the term to the right is the Poisson).

The near equivalence in Eq. (5.32) can be demonstrated by substituting $p = \frac{\lambda}{n}$, then applying conditions 1 through 4 from the preceding list. This proof is left as an exercise for the student (see problem set at the chapter end).

Consider the following example:

Example: On average the hypervariable region of an immunoglobulin is generated with 3 amino acid substitutions per molecule. What is the probability that a hypervariable region is generated with 8 substitutions?

Solution: Since each amino acid is either correct or an error and each amino acid is independently linked into the chain, this is a set of Bernoulli trials. The number of amino acid substitutions has a binomial distribution where n is the number of amino acids in the peptide and p is the probability of error of amino acid assignment. We do not know p or n and so the binomial distribution is not useful to us in this problem. However, the number of errors per molecule is λ and so we do know that $\lambda = 3$. Therefore, the Poisson approximation to the binomial distribution is a reasonable choice for solving this problem. We can write

$$P(8\,\text{amino acid substitutions}) = \frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-3}3^8}{8!} \tag{5.33}$$

Alternatively, we can calculate the probability of there being less than 5 amino acid substitutions in a decapeptide made up from 10 polypeptide chains. Since there are on average 3 errors per molecule, we would expect 30 errors in the decapeptide. We make $\lambda = 30$.

$$P\,(\text{less than 5 errors in decapeptide}) = P\,(0\,\text{errors}) + P(1) + P(2) + P(3) + P(4)$$

$$= e^{-30}\left(1 + 30 + \frac{30^2}{2!} + \frac{30^3}{3!} + \frac{30^4}{4!}\right) \tag{5.34}$$

The Poisson distribution has application to a wide variety of discrete probability problems. An important, widespread application in many biophysical and physiological situations is the use of the Poisson in determining the distribution of arrivals in a given time period. We will consider the problem of arrivals of excitatory potentials at the dendritic tree of a nerve cell. When an excitatory event occurs, the event is recorded and the cell is then immediately ready to receive another input. The average number of excitations in a minute or the rate of excitation is λ. The assumption is that the excitations are independent. We wish to use the Poisson to determine $P(k$ excitations in a minute). In order to use the Poisson distribution we divide the minute into a large number n of small-time subintervals so that at most only one excitatory event can arrive in that time interval. This subdivision makes the arrival or not-arrival of the excitation a discrete Bernoulli trial. The time interval must be chosen so that only one event or no-event occurs. (Note that *this modeling process is an abstraction of a real physical event*. If experimental results do not agree with the mathematical model, this assumption of only one event occurring in a time interval and the independence of events would be good places to start the search for potential theoretical modification). The problem is now appropriately abstracted to a set of Bernoulli trials of subintervals of time n, with a success being the arrival of an excitation. We do not know either n or p (the probability of an excitation arriving), but we do know the rate of excitations arriving in a minute and can use this rate, λ, in the Poisson distribution.

Example: If the basal rate of excitation to a neuron is 3 excitations/s and it will only fire an all-or-none impulse when a rate of 8 excitations/s occur, what will be the probability of the neuron firing "spontaneously" (i.e., what is the probability of 8 excitations/s arriving and coincidently the nerve spontaneously firing)?

Solution:

$$P(8\,\text{excitations in 1 min})\frac{e^{-\lambda}\lambda^k}{k!} = \frac{e^{-3}3^8}{8!} \tag{5.35}$$

Example: Suppose the arrival rate for an excitation is 3/s, what is the probability of at most 5 excitations occurring in 4 s?

Solution: Set $\lambda =$ (the average number of excitations arriving in 4 s) $= 12$, then use the Poisson distribution:

$$P\,(\text{at most 5 excitations in 4 s}) = P(0) + P(1) + P(2) + P(3) + P(4)$$

$$= e^{-12}\left(1 + 12 + \frac{12^2}{2!} + \frac{12^3}{3!} + \frac{12^4}{4!}\right) \tag{5.36}$$

It is often important to have a sense of the level of probability at which we can expect an event to happen. This is called *expectation* and the results of expectation in Bernoulli trials is summarized here:

1) The number of successes in n trials can be expected to be np.
2) The number of Bernoulli trials expected to get a first success is $1/p$.
3) The expected number of character Y objects drawn in n trials is n times the *percentage of character Y* objects in the box.

5.6 Continuous Probability Is Represented as a Density of Likelihood Rather Than by Counting Events

Discrete probability methods are concerned with problems in which a finite or *countably* infinite number of values can be numerically taken by the outcome of the experiment. The numerical outcome of an experiment is called the *random variable* and we have to this point been reviewing discrete random variables. In many physical problems, however, the random variables can take on an infinite number of values, such as any value within an interval. These problems are the concern of continuous probability studies and require different mathematical methods than the discrete techniques used so far. In continuous problems, probabilities of events are assigned by the use of *probability densities*. The probability density function of x, which is continuous, is (x).

Densities are intensive variables that describe a system and relate the number of something with respect to a particular dimension. In chemistry we use the idea of density when describing the number of grams (the something) found in a specific volume (the dimensional view). If a density is constant, then the total amount of the something is simply the length (or equivalent dimensional measure, i.e., area or volume, respectively, in two- and three-dimensional system) multiplied times the density. Densities are often not constant over the interval of concern, however, and then calculus must be used to find the total density over an interval through integration of all of the infinitesimally small densities. Probability densities are not the probability of an event. In the case of discrete problems, the random variable took on a limited number of values and one unit of probability was split into chunks that were assigned to the possible values. When the random variables take on a continuum of value, the probability unit is continuously divided among each point within an interval and the probability assigned to a specific point will be 0 while the probability of the interval will be found by integrating the probability density.

5.6.1 Some Mathematical Properties of Probability Density Functions

Probabilities are never negative and the total probability is always 1, $P(-\infty < X < \infty) = 1$. The density functions are thus limited. They never go below the x-axis, the total area of the function in the interval is 1, and at ∞ and $-\infty$ the functions have always returned to the x-axis. Figure 5.3 shows their form. Density functions are not probabilities. This is because $f(x)$ has the units of probability per unit length (or dimension). Therefore, the probabilities are given by

$$f(x)dx = P(X \approx x) \tag{5.37}$$

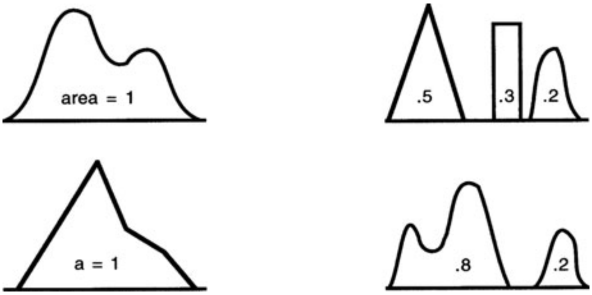


Fig. 5.3 The form of probability density functions

In addition to the density function $f(x)$, which can be used to find the probability of a quantity that has a continuous range of values, another function called a *distribution function* $F(x)$ can also be helpful at times. The distribution function gives a quantity that represents the accumulation of probability and is sometimes called the cumulative distribution function (cdf). All random variables have a distribution function (i.e., the function applies equally to discrete and continuous random variables). The form of cdf is

$$F(x) = P(X \leq x) \tag{5.38}$$

In the cases of continuous variables, the distribution function and the density function are linked by Eq. (5.38) in a manner with significant physical importance:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)dx \tag{5.39}$$

or graphically, in Fig. 5.4 where it is clear that $F(x)$ is the area under the curve and represents the accumulated probability of x.

In the case of a discrete random variable where X can be 1, 2, 3 and the probabilities of these values are $P(X = 1) = 0.7$, $P(X = 2) = 0.2$, $P(X = 3) = 0.1$. The

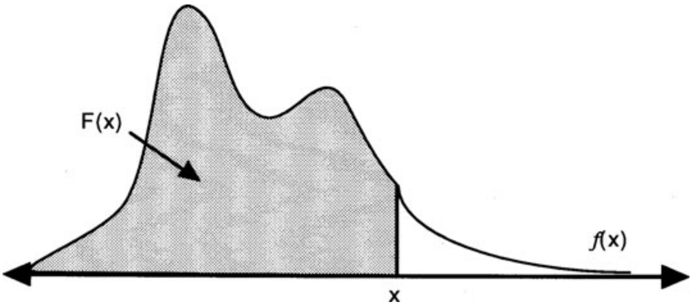


Fig. 5.4 The graphical form of Eq. (5.39)

probability function for X (sometimes called the *discrete density function* of X) is therefore

$$p(x) = \begin{cases} 0.7 if\ x = 1 \\ 0.2 if\ x = 2 \\ 0.1 if\ x = 3 \end{cases} \tag{5.40}$$

which is shown graphically in Fig. 5.5:

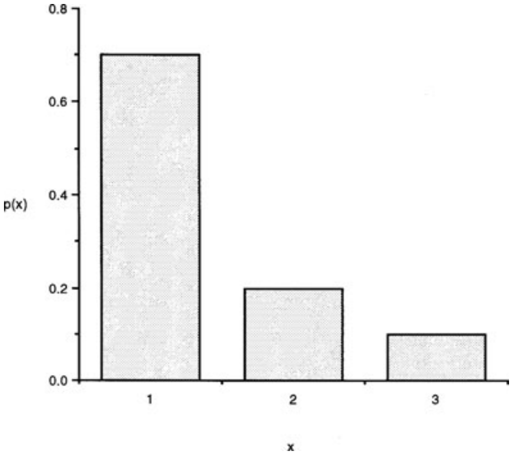


Fig. 5.5 The graphical form of Eq. (5.40)

We will soon discuss functions of this type where the area is concentrated at a point rather than spread out over an interval. Such functions are named *delta functions*. Thus Eq. (5.40) can be written alternatively as

$$0.5\partial(x-1) + 0.2\partial(x-2) + 0.3\partial(x-3) \tag{5.41}$$

If $F(x)$ is the distribution function of X, it represents the cumulative probability of X, which takes the form shown in Fig. 5.6. Here the cumulative probability is 0 until x becomes 1 and then it jumps to 0.7; it stays at 0.7 until x equals 2 at which time

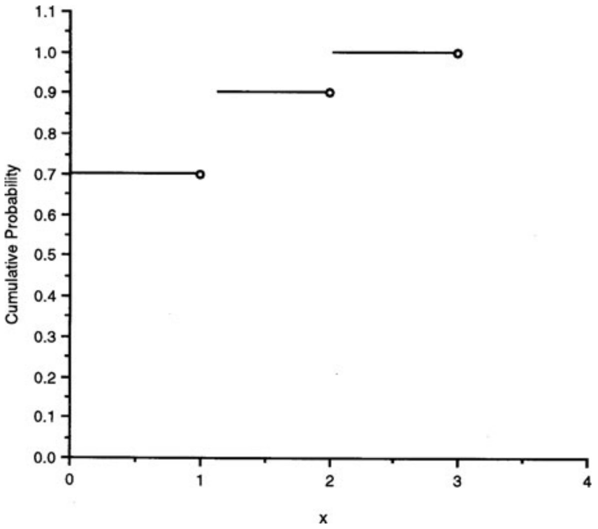


Fig. 5.6 The step function

it jumps to 0.9; again, it stays at 0.9 until x reaches 3 at which time the probability becomes 1 and remains there unchanged. Therefore, we can see that the distribution function of a discrete random variable is a *step function* that rises from 0 to 1 with jumps at each of the possible values of X.

We can summarize the properties of distribution functions as follows, if X is a random variable with distribution function $F(X)$:

1) F is non-decreasing, which means it can increase or remain constant but it cannot go down (or have a negative slope at any point)
2) $F(-\infty) = 0$
3) $F(\infty) =1$

Distribution functions can be used to find probabilities. If there are jumps in the function, then at the jump $x = x_0$, the probability increases, $P(X{=}x_0)$. When the function is represented by horizontal line, probability is not accumulating and the density must be 0 and events have a probability of 0.

All random variables have probability functions but not all random variables have density functions. If a random variable has a density function, it is continuous. Continuous random variables have the property that $P(X = x) = 0$ which means that the distribution function cannot jump. For a continuous function, the distribution function is the area under the curve of $f(x)$:

$$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(x)\, dx \tag{5.42}$$

We will consider several important distributions, the exponential, Gaussian (or normal or error), and the Boltzmann distributions.

5.6.2 The Exponential Density Function Is Useful for Lifetime Analysis

The exponential density can be written for λ > 0 and $x = 0$:

$$f(x) = \lambda e^{-\lambda x} \tag{5.43}$$

The exponential density has important use in determining waiting times and lifetimes. λ is the arrival rate or the average number of arrivals per unit time, and X is the waiting time or the waiting time between arrivals. (Letting $F(x)$ be the distribution function of X and then getting $f(x)$ is easier). Waiting times cannot be negative so $F(x)$ equals 0 if x is less than or equal to 0 and for x is greater than or equal to 0:

$$F(x) = P(X \leq x) = 1 - P(X > x) \tag{5.44}$$

If we have waited more than x time for the next arrival, no particles arrived during x and therefore

$$F(x) = 1 - P\,(\text{no arrivals in}\,x\,\text{min}) \tag{5.45}$$

The number of arrivals in x min is a discrete random variable with a Poisson distribution with the form (the parameter λ is the average number of arrivals in x units of time):

$$F(x) = 1 - \frac{e^{-\lambda x}(\lambda x)^0}{0!} = 1 - e^{-\lambda x}\,\text{for}\,x \geq 0 \tag{5.46}$$

Differentiating $F(x)$ gives

$$f(x) = F'(x) = \lambda e^{-\lambda x}\,\text{for}\,x \geq 0 \tag{5.47}$$

This problem has an exponential distribution. λ is the arrival rate and 1\λ is the average waiting time. Thus, if the particles arrive independently and X is the waiting time for the next arrival, then X has an exponential distribution with λ as described.

The lifetime analysis is quite similar. Here the age of a particle has no effect on whether it exists or dies. X is the lifetime of the particle. If the particle eventually is eliminated and is replaced by another particle that then eventually is eliminated and is replaced, the average lifetime is the time waiting for elimination. By making the hypothesis that the likelihood of elimination is the same for the particle regardless of its age, means that the replacement-eliminations are independent. Thus X is the waiting time for replacement-eliminations and has an exponential distribution with

$\lambda = 1/\text{average lifetime}$. An important feature of exponential distributions is that they are memoryless (in fact they are the only memoryless continuous function). This means that the conditional probability that you must wait x more hours for an event after having already waited y hours is equal to the unconditional probability of having waited x hours initially. If the lifetime of a particle has an exponential distribution, then the probability that the particle will last x hours is the same for all particles regardless of their age. Thus, if in an experiment, entities were to show lifetimes with an exponential distribution, we would infer that these entities would possess the property of being memoryless with respect to their formation–elimination.

5.6.3 The Gaussian Distribution Is a Bell-Shaped Curve

The normal or Gaussian distribution is seen so frequently in physical problems as the form in which many random variables seem to take that in some ways it has begun to suffer the contempt reserved for familiar axiomatic knowledge. Because physical measurements are always subject to error and the distribution of the error is given by the Gaussian, this is a fundamental distribution. In a large enough population, the Gaussian describes the distributions of height, weight, test scores, and neuromuscular response time to name just a few. The bell-shaped curve associated with this function comes from the shape of the graph of $y = e^{-x^2}$. The area under this curve is generally called the *error function* (erf) and has the form:

$$\mathrm{erf}(x) = \frac{2}{\sqrt{2}} \int_0^x e^{-t^2} dt \tag{5.48}$$

The Gaussian distribution function is a closely related function to the erf and is given the form:

$$P(-\infty, x) = \frac{1}{\sqrt{2\_}} \int_0^x e^{-t^2} dt = \frac{1}{2} + \frac{1}{2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{5.49}$$

or for the interval between 0 and x:

$$P(0, x) = \frac{1}{2}\mathrm{erf}\left(\frac{x}{\sqrt{2}}\right) \tag{5.50}$$

The Gaussian distribution is actually an approximation to the binomial distribution for large n and for a continuous random variable. For large n, where the discrete n of multiple Bernoulli trials are so numerous that they are closely packed together, it is easy to imagine that the continuous nature of the Gaussian closely approximates the binomial distribution quite satisfactorily. It is somewhat pleasantly surprising, however, that the normal error curve is also quite good in approximating even small values of n. The probability of getting between 45 and 55 heads in 100 coin tosses is given exactly by the area under a binomial $f(x)$ curve but for a large n, it is easier to integrate the area under the normal approximation curve (another name for the Gaussian):

$$P(x_1 \leq x \leq x_2) \approx \frac{1}{\sigma\sqrt{2\pi}} \int_{x_1}^{x_2} e^{-(x-\mu)^2/2\sigma^2} dx \tag{5.51}$$

It is clear by inspection that the ratio of the two sides tends toward 1 as $n \to \infty$. This statement is named the *Laplace–De Moivre* limit theorem. We have used the density function of the Gaussian with parameters μ and σ which are the mean and the standard deviation (σ^2 is the variance) of the random variable. In the terms of the binomial distribution, $\sigma = npq$ and $\mu = np$ where n is the number of trials, p is the probability of success, and q is the probability of failure for the trial. The mean is the expected value in a set of trials and in the Gaussian distribution is the point of maximum density with the density then symmetrically spread out around the mean with the measure of spread around the mean given by σ^2, the variance (Fig. 5.7). The variance controls the width of the Gaussian distribution. The smaller σ is, the narrower the distribution becomes. In the limit $\sigma \to 0$, Eq. (5.51) becomes a *delta* function, in which all the probability becomes concentrated at a single point.

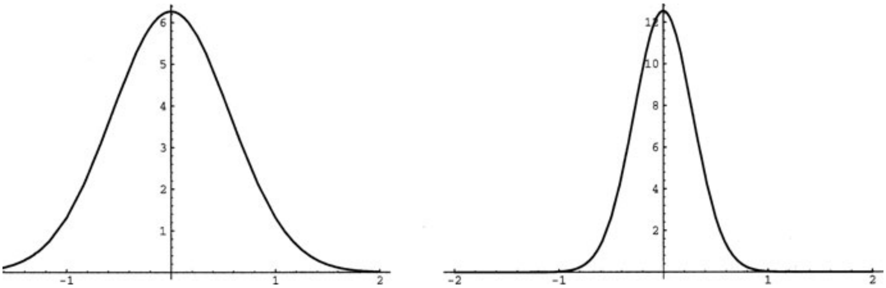


Fig. 5.7 The normal distribution. The distribution on the *left* has a larger σ compared to the curve on the *right*

When $\mu = 0$ and $\sigma = 1$, the normal distribution is called the *unit* or *standard normal*. It is these standard values that are tabulated in tables. Alternatively, the standard normal density function ($\phi(t)$) can be found by calculation using

$$\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \tag{5.52}$$

Generally we will find few physical problems in which the parameters come in standard form, but since the distribution is the same shape no matter where the mean falls for a given variance, we can convert between any case and the standard tables by recognizing the new axis $(x - \mu)$ as a new variable. Making the x-axis scale in terms of the standard deviation, σ, we write

$$t = \frac{x-\mu}{\sigma} \tag{5.53}$$

and

$$f(x) = \frac{1}{\sigma}\phi\ (t) \tag{5.54}$$

5.6.4 Stirling's Formula Can Be Used to Approximate the Factorials of Large Numbers

Factorials of very large numbers are frequently encountered in biophysical chemical problems. Consider for a moment the factorial of Avogadro's number! Numbers of such great magnitude are quite difficult to calculate and manipulate. My handy scientific calculator hits its maximum at 69! as the exponent of 70! goes over 100. Calculating by hand is no rational substitute. An approximation of $N!$ for large N, called Stirling's approximation, is often used. This is an asymptotic approximation which is an approximation to a function that becomes more accurate as the magnitude of the function's argument increases. Stirling's approximation is actually an approximation to ln $N!$, since $N!$ is a product and ln $N!$ is a sum which is actually easier to work with. The result of the approximation is

$$\ln N! = N \ln N - N \tag{5.55}$$

The derivation of Eq. (5.55) proceeds as follows:

Since

$$N! = N(N-1)(N-2)\cdots(3)(2)(1) \tag{5.56}$$

then:

$$\ln N! = \sum_{m=1}^{N} \ln m \tag{5.57}$$

If the graph of ln x versus x is considered, we can see why the approximation is reasonably made that Eq. (5.57) can be replaced by an integral:

$$\ln\ N! = \sum_{m=1}^{N} \ln\ m \approx \int_{1}^{N} \ln\ x\, dx = N\ \ln\ N - N \tag{5.58}$$

The sum of the area of the rectangles up to x is equal to N is ln $N!$. If a continuous curve of ln x is superimposed on the rectangles, it can be seen that as x grows larger, the curve more precisely approximates and smoothly encloses the sum of

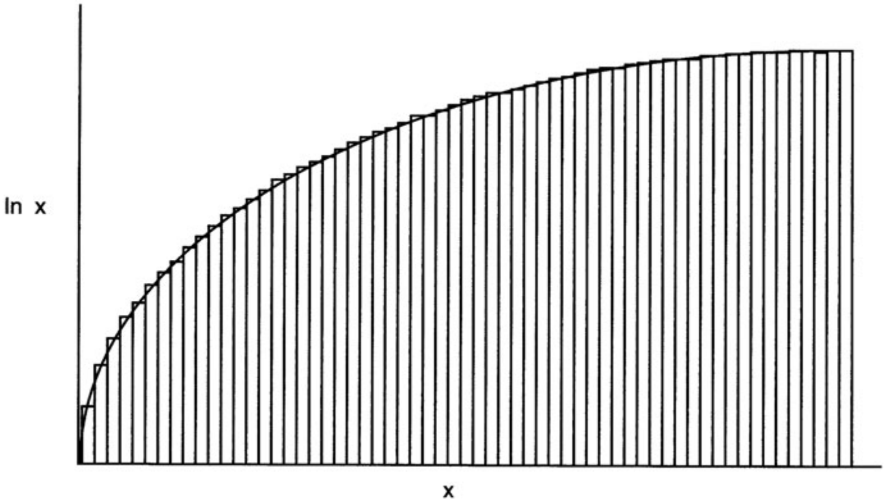


Fig. 5.8 Graphical demonstration of the validity of the Stirling approximation

the rectangles (Fig. 5.8). As x increases, it is reasonable to say that the area under the rectangles can be approximated by the integral of the curve, ln x. Although initially the area under the curve ln x poorly approximates the area under the rectangles, as x becomes large (and this is the basis of an asymptotic approximation), this poorly matched area will make a very negligible contribution to the total area. Thus Eq. (5.58) is reasonable.

5.6.5 The Boltzmann Distribution Finds the Most Probable Distribution of Particles in Thermal Equilibrium

An important question in many physical studies is: Given a system at thermal equilibrium, what is the most probable distribution of the particles in that system. A system that contains a large number of similar physical entities at equilibrium must be arranged so that the entities can exchange energy with one another. As a group they will have a certain average energy, but at any given moment the energy content of each entity will fluctuate because of the exchange process. Some will have more than the average energy and some will have less. These energies will be distributed according to a specific probability distribution whose form will be determined by the equilibrium temperature. The average energy value of each entity will be a function of the probability distribution and thus linked to the temperature.

The system for which a most probable distribution of states is sought is generally constrained in several ways. First of all, it is an isolated system, so its energy, U, is constant. Furthermore, in an isolated system at equilibrium, the total number of

Table 5.1 Arrangement of four entities that leads to the form of the Boltzmann distribution

	$E = 0\Delta E$	$E = 1\Delta E$	$E = 2\Delta E$	$E = 3\Delta E$	$E = 4\Delta E$	# of duplicates *(distinguishable)*	P_i
$i = 1$	3	0	0	1	0	4	$\frac{4}{20}$
$i = 2$	2	1	1	0	0	12	$\frac{12}{20}$
$i = 3$	1	3	0	0	0	4	$\frac{4}{20}$

particles, N, may also be considered constant. We can deduce the form of the distribution by consideration of a simple case. Our simple case involves four entities that each may have an energy of either $0\Delta E$, $1\Delta E$, $2\Delta E$, $3\Delta E$, or $4\Delta E$. These entities share a total of $3 \times \Delta \mathrm{E}$ among themselves and this energy can be exchanged freely. The energy can be divided among the four entities as shown in Table 5.1. Because we treat each of the identical entities as if they are distinguishable, the arrangements according to each energy distribution are listed. Note that the identical entities are treated as distinguishable except for rearrangements within the same energy state. This is the central characterization of the Boltzmann distribution. So the probability of an arrangement will be the permutations (rearrangements) of the number of entities corrected for the number of permutations of entities in the same energy division. In our case there are 4! permutations reduced by 3! for states $i = 1$ and $i = 3$ and by 2! for $i = 2$:

$$\text{State1} \qquad \frac{4!}{3!} = 4$$

$$\text{State2} \qquad \frac{4!}{2!} = 12$$

$$\text{State3} \qquad \frac{4!}{3!} = 4$$

The other assumption in the Boltzmann distribution is that all of the divisions of energy are equally probable and the relative probability P_i is the number of rearrangements divided by the total possible number of these divisions (20). Now we calculate $n(E)$ the probable number of entities in each energy state, E. This is determined by considering the number of entities in each state times the relative probability P_i for that division.

For

$$E = 0; \left(3 \times \frac{4}{20}\right) + \left(2 \times \frac{12}{20}\right) + \left(1 \times \frac{4}{20}\right) = \frac{40}{20}$$

$$E=1;\left(0\times\frac{4}{20}\right)+\left(1\times\frac{12}{20}\right)+\left(3\times\frac{4}{20}\right)=\frac{24}{20}$$

$$E=2;\left(0\times\frac{4}{20}\right)+\left(1\times\frac{12}{20}\right)+\left(0\times\frac{4}{20}\right)=\frac{12}{20}$$

$$E=3;\left(1\times\frac{4}{20}\right)+\left(0\times\frac{12}{20}\right)+\left(0\times\frac{4}{20}\right)=\frac{4}{20}$$

$$E=4;\left(0\times\frac{4}{20}\right)+\left(0\times\frac{12}{20}\right)+\left(0\times\frac{4}{20}\right)=\frac{0}{20}$$

If $n(E)$ is plotted against E (Fig. 5.9), the points fall very closely to a curve of a decreasing exponential function:

$$n(E)=Ae^{\frac{-E}{E_0}} \tag{5.59}$$

If the interval between the energy states is now allowed to shrink while increasing the number of allowed states, all of the points on the plot will converge to fall directly on the continuous function line graphed in Fig. 5.9. This is the probability function of the Boltzmann distribution. If the law of equipartition of energy is applied, E_0 will be equal to the average energy of an entity and thus is equal to kT. This leads to the Boltzmann distribution:

$$n(E)=Ae^{\frac{-E}{kT}} \tag{5.60}$$

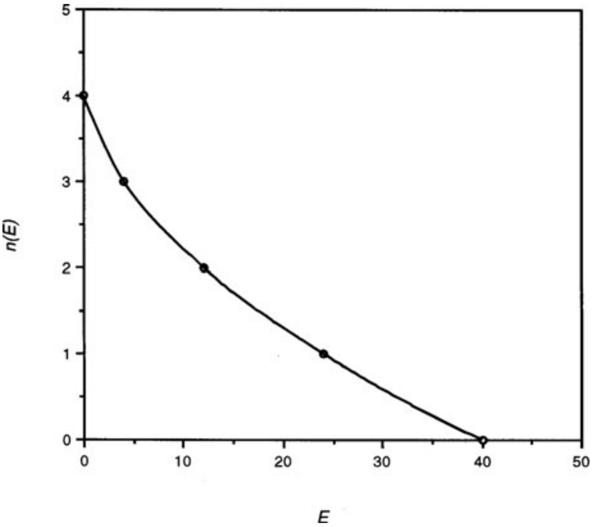


Fig. 5.9 The Boltzmann distribution function (drawn as a *solid line*) is closely approximated by the simple case study (*circles*) described in the text

Further Reading

Ash C. (1993) *The Probability Tutoring Book*. IEEE Press, New York.

Berg, H.C. (1993) *Random Walks in Biology*, Revised edition. Princeton University Press, Princeton, NJ.

Dill K.A. and Bromberg S. (2003) *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology*. Garland Science, New York.

Ott R.L. and Longnecker M.T. (2008) *An Introduction to Statistical Methods and Data Analysis*, 6th edition. Duxbury Press, N. Scituate, MA.

Problem Sets

1. Would a base 4 or 5 system of numbering be beneficial in devising mathematical methods for handling problems dealing with information transfer in DNA and RNA research? Discuss why or why not.
2. What is the total number of configurations possible in a peptide chain composed of 150 amino acids each of which can take one of four conformations?
3. Show that $\binom{n}{r} = \binom{n}{n-r}$.
4. Show that $\binom{n}{1} = n$.
5. Show that $\binom{n}{n} = 0$.
6. Graphically show the distribution function of a uniformly distributed random variable.
7. You have 20 amino acids available. How many pentapeptides can be made if an amino acid is not replaced after being used?
8. How many permutations are there of 10 nucleotides that form a decapolynucleotide?
9. Evaluate the following combinations:

   (a) $\binom{40}{6}$ (b) $\binom{35}{5}$

   (c) $\binom{20}{6}$ (d) $\binom{75}{20}$
10. Find the approximate probability of getting fewer than 52 heads in 100 tosses of a coin.
11. You have 20 adenosines, 50 thymidines, 15 guanosines, and 35 cytosines. Pick 15 sequential nucleosides. What is the probability of getting 5 thymidine, 5 adenosine, 5 cystosine and 1 guanosine?
12. Evaluate: $\ln 1.8 \times 10^{18}!$
13. Prove that the binomial distribution and the Poisson distribution are nearly equivalent.

$$\binom{n}{k} p^k(1-p)^{n-k} = \frac{n(n-1)\cdots(n-k+1)}{k!} p^k(1-p)^{n-k}$$

14. The exponential distribution is a special case of the more general *gamma distribution.*

$$f(x) = \frac{l^n e^{-\lambda x} x^{n-1}}{(n-1)!}$$

The gamma distribution has two parameters, n and λ. Show that the exponential distribution is the special case of the gamma function when $n = 1$.

Part II
Foundations

Chapter 6
Energy and Force – The Prime Observables

Part of a scientist's search for simplicity is the search for economy of concepts. A physical concept must not be merely quantitatively definable. . .it must bring something of special value to the description of nature. By every criterion. . .the concept of energy deserves attention.

From the nuclear energy of the sun's interior to the work performed by a man's muscle runs a long and elaborate path of energy transformations through the sciences of astronomy, physics, chemistry, geology and biology. Because of its conservation at every step along the way, energy is an unbroken thread tying these disciplines together.

WF Ford, 1968

Contents

6.1 Experimental Models Are a Careful Abstraction of Either Descriptive or Explanatory Models

The idea of abstractions as *practical* simplifications of systems has been introduced because it is fundamentally useful in the day-to-day laboratory experience of the

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_6, © Springer Science+Business Media, LLC 2010

scientist. By using different levels of abstraction while making observations on a system of interest, we can invoke different techniques. We often find ways to imply information that might not otherwise be accessible to direct measurement. When we examine a system we decide at what level our detailed knowledge of the system will be needed and choose an abstraction appropriate to that level of detail, much like we might choose the magnification on a microscope. They are simplifications and do not operate out of context. Abstractions are attempts to generalize about systems whose actual behavior may be different from that predicted by the abstraction. Detecting these gaps between real and predicted behavior *is a central role of experimental science*. We can formulate an image of a natural system with its peaks and valleys (represented by functions); then, choosing an abstraction, we predict a behavior of the system. The behavior of the real system is then measured and mapped to confirm or falsify the prediction. This map of the natural system's state space gives us the information we need to describe the system and to predict its future action. Several of the most important maps that we can construct for a biological system are maps of its energy at each point of the state space and of the forces found acting at each of these points. These energy and force mappings are a unifying idea that is used in every level of abstraction.

6.2 Potential Energy Surfaces Are Tools that Help Find Structure Through the Measurement of Energy

The primary goal of biophysical chemistry is to describe a system by relating its function and its structure. It is essentially a discipline devoted to determining the structure and action of a system. We have already seen that a systemic approach gives us tools to describe the observables of any system whether the system is a molecule, an organelle, a cell, an organism, or a system of organisms. The system is a structure defined by knowledge of the position of its elements in space and time. To lend concreteness to this discussion, consider a protein and its tertiary structure (see Fig. 25.8). A single instantaneous snapshot of a system can describe the relative positions of each element without considering any temporal dimension. Such a time-independent description provides a static structure. When we think in these terms, space and time become separated. While it is often convenient and practical to make this separation, spatial and temporal dimensions are not, in fact, separate. In a natural system, space and time move forward together in a four-dimensional state space and the complete description of our protein's structure must set forth the relative positions of each atom at a given time. This is important because if we take multiple snapshots of our protein over a period of time, we discover that there is some movement of each atom from time point to time point. In other words, the elements of the protein are in motion relative to one another. Over a somewhat extended period of time we will find one of two general cases.

(1) The relative motion of the atoms in the protein will either continue to move toward each other or away from one another. If they continue to move apart, the molecule will reach a point where it is no longer able to hold together and the protein denatures and the structure falls apart. Alternatively the atoms can keep moving toward each other until they smash into one another again irreversibly changing the initial observed structure of the protein. We can relate this process to our discussion of equivalent systems in Chapter 4. Because we are describing our protein by defining its relative atomic positions, it is easy to see that a form is being mapped onto a geometric coordinate system. As the atoms move closer or farther apart there comes a point at which no transform of the coordinate space will allow us to recognize the similarity of the preceding state to the following state. The point at which the molecule can no longer be distorted but rather breaks, is a bifurcation in state space, and a new, non-equivalent species is formed.

(2) An alternative to seeing the protein denature is the finding that the motion of the atoms shows that for a certain time the atoms move apart, but are then seen to slow, stop, and reverse direction. Thereafter the atoms are seen to move toward each other again, typically slowing and reversing direction before smashing into one another. Each atom acts as if it is connected by a spring and vibrates back and forth. Over time, this molecule will have an average structure that can be found by adding up all the positions of each atom as it vibrates along its “spring” line. This molecule viewed over time has an average or equilibrium structure which is a summation of many time-independent snapshots, but which represents a highly dynamical system in space-time.

Let us consider several points about these examples. First we have described a molecular system by defining the coordinates of each atom in the system. We have implicitly treated the atoms in the molecule as bound together in some fashion. Next we have acknowledged that the positions of each atom in the molecule vary over time. In one case we have argued that the positions of the atoms oscillate around an average value, resulting in a net or average structure, the equilibrium structure of the molecule. In the other case we noted that if the motion of the atoms is such that they do not oscillate but rather move apart, at some point the molecule passes a bifurcation point and becomes a new category molecule (i.e., denatured, broken bonded, etc.). We mentioned the opposite case of smashing all of the atoms together, but this is probably not a case of practical biological interest.

What we have accomplished is not trivial. We have a way of describing an average structure based on positions in coordinate space. Moreover, since the atoms of the system are moving around an equilibrium position they may be described as having both kinetic energy associated with their motion and potential energy associated with their position. There are forces that deform and restore the atomic particles from and to their equilibrium positions and these forces may be intrinsic (forming bonds) and extrinsic (representing the environmental forces). We can make a map of the state space of this molecule and relate each coordinate point to the kinetic and/or potential energy and thus the total energy. The surface made by such a map is an energy, or more commonly a potential energy surface of the state space. The forces acting at each point of the coordinate space are derivatives of the potential energy with respect to position. A potential energy surface can be related to a surface that links the positions with the magnitude and directions of the forces: a vector field.

The central concern with physical studies is to know where things are in space, which is a function of the coordinate system and is thus often written $f(x, y, z)$. We also wish to know how the things are moving, which is measured by their momentum. This is just another way of saying that we wish to know the potential energy (dependent on position in a potential field) and kinetic energy of the objects in a state space. Why do we care about such things? The answer is evident when we consider the temporality of the space-time dimension. There is distinct value in being able to anticipate where a thing may be in the future and what it will look like when it gets there. Implicit in this is the question how will it be moving when it reaches its future engagement? Consider the first example given earlier: if the positions and momenta of the atoms in our protein are such that continued motion in a particular direction is going to lead to bond breaking or denaturation of the protein, then our physical knowledge allows us to anticipate the catastrophic events that are chemical in nature, i.e., the conversion of reactants to products. Thus physical motion over time is the essence of chemical and biological reactions. Using potential energy surfaces, knowing physical coordinates, and having a knowledge of the linkages between the forces in a state space provide us with the tools (observables) necessary to anticipate, participate in, and guide future events. The value and interest in biophysical chemistry should be self-evident.

Underlying this is a very important idea: We *define a conservative system as one in which the energy of a point in state space is related to its position*. Gravity and electrical potential are just two examples of these fundamental conservative forces. In fact, non-conservative forces such as friction are really only artifacts of mechanical abstractions. At the molecular level we tend to be interested in all of the forces contributing to a system and so, in fact, all forces are conservative. Thus making a distinction between conservative and non-conservative forces is artificial and we will not treat non-conservative forces in any significant degree. Consider the potential energy surface for a system. It relates position to energy and as such relates structure and energy. The positions that a system is likely to occupy and thus be recognized as “in” occur when a minimum value on the potential energy surface is found. Because systems tend to oscillate around the minimum value on a surface, the minima are described by curved (i.e., continuous) functions. We can draw the potential energy surface by moving the elements of the state around to different points and determining the different energies and forces at each point surrounding the minima. (We should recognize that the potential energy surfaces have maxima as well and since most potential energy surfaces of interest are complex, the maxima that separate adjacent minima are called *saddle points*.) Energy surfaces are determined as follows: we move physical elements to different points in state space and measure the energy, or we do a virtual experiment and calculate the expected energies at each coordinate point in state space. (The latter is the essence of computational chemistry.) If the elements in state space just jiggle around the equilibrium points described by the minima, the equilibrium structure described above will be found. However, elements in state space can move over the potential energy surface if they gain enough kinetic energy so that they can overcome the potential energy barriers (maxima) that separate the energy wells. The movement of a system across its potential energy surface depends on the kinetic energy distribution to the elements. Knowing the total energy of a system in relation to its potential energy surface allows us to anticipate the overall ability of the system to assume variety of accessible states. We can anticipate that the net structure of a complex system, such as a protein, will be the sum of a series of structures that can be plausibly determined. The possible structures will be dependent on the potential energy barriers between the different structures and the amount of kinetic energy available to move the system between states. The ability to change between potential energy wells or states will depend over time on whether enough kinetic energy is available to make the jump (i.e., the activation energy) to a new well and whether the well being jumped into is occupied or vacant. If the potential well is already occupied, double occupancy will lead to the case of smashing objects together which is generally energetically unheard of in chemical systems. It does not happen. The measured rate of change of state across a potential energy surface is the time dependency of the reaction coordinates and is the focus of kinetic studies. Thus the overall state of a system can be plotted on a potential energy surface. The observed state of the system is a function of both the overall favored potential energy state and the kinetic likelihood that the system can get there. Many metastable states exist in nature in which the preferred potential energy state is not attained, because the kinetics of getting into the state are either blocked or so slow that the system is left “frozen” along the way.

6.3 Conservative Systems Find Maximal Choice by Balancing Kinetic and Potential Energies Over Time

Two very fundamental principles relate the observed aspects of a system to the balance between kinetic and potential energy and the time it takes for a system to find the balance. These principles are as follows:

(1) Fermat’s principle of least time.
(2) Hamilton’s principle of least action.

These principles relate the physical nature of a system to time and specify four-dimensional constants (space-time). We will come to them in due time in the following pages as we explore the tools used to characterize the energy of systems, namely classical mechanics, thermodynamics, and quantum mechanics. Here we state a foundation relationship that is basic to almost everything that follows in this book: The law of conservation of energy states that the total energy of a system is fixed and is equal to the sum of the kinetic and potential energy in the system. The forms of kinetic and potential energy may change and portions of energy may be partitioned in any fashion between these forms but the total energy of the system does not change. Thus

$$E_{\text{total}} = \text{PE} + \text{KE}. \tag{6.1}$$

Conservation laws are very important and in fact appear to be the most fundamental statements of natural law. An important philosophical character of conservation laws is that they are laws of prohibition *not* permission. A law of prohibition allows anything to happen as long as it respects the boundaries of prohibited action. This is a system of probability, within certain constraints there are many possibilities. Deterministic laws are permissive, i.e., they state what can and should happen. These permissive laws are laws of certainty and restrict choice. As we will see in our study, the laws of nature are fundamentally probabilistic not deterministic. Thus the laws of conservation appear to be consistent in character with these fundamentals. There are six laws of conservation:

(1) Mass and energy
(2) linear momentum
(3) angular momentum
(4) charge
(5) lepton number
(6) baryon number

The first three of these laws mean that the results of an experiment are independent of the experiment’s position in space-time. The last three are book-keeping, building-block counting, rules for subatomic particles. The quantum mechanical proposition that things do not depend on absolute time can be combined with the other rules of quantum mechanics to produce the law of conservation of mass and energy. The second and third laws are similar. The second law assures us that the position in space makes no difference for the outcome of the experiment; the third says that we can rotate a system in space with no change in the observables of the system. These quantities are all conserved because space (actually space-time) is isotropic (the same in every direction) and homogeneous (the same everywhere). This *invariance* from which conservation is derived is a consequence of *symmetry*.

Although each conservation law leads to maximum possibilities, in any situation *all* of the laws must be obeyed. This can make the conservation laws very restrictive sometimes allowing only a single possibility. We will discuss the last three laws here only to point out that the laws of conservation and especially the law of conservation of charge are responsible for the stability of the electron and therefore for the stability of the whole of chemistry and biological life. The electron is the lightest charged particle and cannot decay. Only the neutrino and photon are lighter than the electron and they are uncharged. If an electron were to decay, charge would not be conserved. If the law was not completely correct the electron would have a finite lifetime. Experiment has shown that the lifetime of an electron is $>10^{21}$ years and so, to a very significant degree of confidence, conservation of charge must be valid.

6.4 Forces in Biological Systems Do the Work That Influences Structure and Function

Energy, force, and work are central to the structure and function of systems of physical interest and our knowledge of a system depends on interaction with these functions in order to make an observation. The way that a particle experiences the existence of another near-by particle is wholly dependent on an interactional force exerted between the particles. If there is no force there can be no response of the other particle. As observers of a system we must interact with it to have any knowledge about it. As we have just seen, having knowledge of a system's energy state can be very useful when we want to understand the system's structural and kinetic aspects. This is a central tenet throughout this book. We will see that the important forces that act in biological systems are largely mechanical, which includes mechanical actions induced by thermally induced random motion, and electrical, which includes the formation of chemical bonds, and therefore underlie chemical reactions.

6.4.1 The Concept of Forces and Fields Is Derived from Newton's Laws of Motion

Isaac Newton described his laws of motion first in 1687 in the monumental work, *Philosophie Naturalis Principia Mathematica*. These laws form the core of classical mechanics which is a theory of motion based on the ideas of mass and force and the formal rules that relate these quantities to the properties of motion, position, velocity and acceleration. Newton's laws, including the law of gravitation, which was presented in the *Principia* accurately describe the way bodies move in virtually all observable situations except very small particles and motion at speeds that approach the speed of light. The laws of motion describing these mechanical systems are quantum mechanics and relativity and have been shown to be the general laws of mechanics. Newtonian mechanics is a special case. In so many cases, from the movement of cells to the motion of planets and especially on the human scale, Newtonian mechanics have been shown to be accurate and valuable so that a sound understanding and appreciation for these laws is still a requirement for the modern scientist.

The first law describes *inertia* and says that a body, in an inertial frame of reference, with no external forces acting on it will either remain at rest or remain at a constant velocity.

The second law relates forces acting on a body to its velocity and acceleration. It says that the rate of change in momentum of a body with respect to time is equal to the resultant external force acting on the body. This leads to the famous mathematical expression of this law:

$$\sum \mathbf{F} = \frac{d\mathbf{p}}{dt} \tag{6.2}$$

Momentum, $\mathbf{p}$, is defined as mass times velocity, $\mathbf{p} = mv$. For a body of constant mass we can write as

$$\sum \mathbf{F} = m\frac{dv}{dt} = m\mathbf{a} \tag{6.3}$$

The third law is the law of action and reaction and states that forces occur in equal but opposite pairs. Thus if a body X exerts a force $\mathbf{F}_{az}$ on a body Z, Z also exerts an equal but opposite force $\mathbf{F}_{zx}$ on X.

Finally forces are treated as vectors.

6.4.2 Force and Mass Are Related Through Acceleration

Newton's laws give us an operational definition for the ideas of *force* and *mass*, which can be quite difficult to define. A net external force is defined to be present when a body at rest moves or a body of constant velocity changes its velocity vector. The first law thus gives a way of knowing if a force is present. The second law gives us the tools to find and define the forces acting in the inertial frame of reference. The concept of mass is defined within the framework of the second law. It helps us quantitate the following experience: If we exert an equal pull on two objects in identical circumstances (i.e., apply an equal pulling force to two boxes lying on a frictionless plane) and one box moves more quickly in a defined time interval than the other, we can know that a more rapid acceleration was induced in the less *massive* body than in the other. The concept of mass can be quantified by measuring the acceleration that a given force (perhaps measured in our experiment by a spring scale) induces in different bodies. The ratio of the mass of two objects is inversely proportional to the accelerations produced by the same force on each body. Thus we can define mass as

$$\frac{m_2}{m_1} = \frac{a_1}{a_2} \tag{6.4}$$

Mass in a Newtonian system is an intrinsic property of an object. The acceleration produced in two bodies is independent of the magnitude and direction of the force. Acceleration is independent of the type of force that induces it. The intrinsic properties of mass allow it to be treated as a scalar quantity obeying the normal rules of arithmetic and algebra.

The units of force, mass, and acceleration are chosen so that the constant of proportionality that is implicit in Eq. (6.3) is unity. Thus a force of 1 unit acting on a mass of 1 kg will produce an acceleration of 1 m/s^2. This unit force is called the *Newton* and has units of

$$1\,\mathrm{kg\,m/s^2} = 1\,\mathrm{N} \tag{6.5}$$

6.4.3 The Principle of Conservation Leads to the Concept of a Force Field

Newton's third law of paired forces acting as equal action–reaction pairs leads to the conclusion that for two objects isolated from their surroundings, experiencing only mutual forces, the sum of their momenta remains constant in time. This means that their momenta are conserved. This can be written mathematically as

$$\mathbf{p}_1 + \mathbf{p}_2 = \text{constant} \tag{6.6}$$

This result is derived as follows: $\mathbf{F}_1$ is the force exerted on one object and $\mathbf{p}_1 = m_1 v_1$ is its momentum. The second law gives $\mathbf{F}_1 = \dfrac{d\mathbf{p}_1}{dt}$. For the second object $\mathbf{F}_2 = \dfrac{d\mathbf{p}_2}{dt}$. Since $\mathbf{F}_1 = -\mathbf{F}_2$:

$$\frac{d\mathbf{p}_1}{dt} = -\frac{d\mathbf{p}_2}{dt} \tag{6.7}$$

or

$$\frac{d\mathbf{p}_1}{dt} + \frac{d\mathbf{p}_2}{dt} = \frac{d}{dt}\left(\mathbf{p}_1 + \mathbf{p}_2\right) = 0 \tag{6.8}$$

Thus the rate of change of the summed momenta is zero and Eq. (6.6) must hold.

The principle of conservation of momentum can be generalized to any system of isolated bodies. The observation that one body gains momentum at the same rate that another body loses momentum is easily appreciated when the bodies are in close proximity or, even better, in contact. However, if they are widely separated such as the Earth and Sun, conservation of momentum implies an instantaneous transmission of momentum across space or *action at a distance*. This idea bothered Newton but he recognized the practical value of the principle of momentum since it allowed him to calculate the correct orbits of the planets from the law of gravitation. He considered this concept as an approximation and at best an operational abstraction. The problem of action at a distance is treated by introducing the concept of a *field*. A field is a condition in space created by the body such as the Earth or Sun, which in the cases of planetary motion create gravitational fields. These fields then interact one with the other and a change in the position of one of the objects leads to a change in the propagated field. The speed of field propagation is limited to the speed of light and so *if* the time of field propagation can be ignored, the action–reaction exchange is effectively instantaneous.

6.4.4 Energy Is a Measure of the Capacity to Do Work

Work in biological systems usually takes one of several forms: mechanical, electrical, chemical potential, and to a lesser degree, pressure–volume work. *Mechanical work* is movement and includes: chromosomal movement in metaphase due to the contraction of microfibrils, ruffling of cell membranes, migration of cells, and muscular contractions that pump blood or move arms and legs. The *chemical potential* causes *diffusion* of chemical species down a concentration gradient. *Electrical work* maintains concentration gradients across membranes, often countering chemical potential work. It removes toxic and waste materials; provides information transfer between neural, neuromuscular, and other cells; and extracts the energy from either the Sun or from reduced substrates of carbon to provide stores of chemical energy. *Pressure–volume work* occurs with the expansion or contraction of a gas and is important in gas-exchanging organs.

All *work* is characterized by the displacement or movement of some object by a force. All forms of work take the general form:

$$\text{Work} = \text{Force} \times \text{Displacement} \tag{6.9}$$

By convention, work done on a system is considered positive. In other words, when the surroundings do work on the system, a positive sign is given to the value. On the other hand, if the system exerts a force that causes a displacement in the surroundings, the work is said to be negative, representing work done on the surroundings.

6.4.4.1 Electrical Work

Electrical work is defined as the movement of charge through an electrical gradient or potential:

$$w_{\text{electrical}} = -EQ \tag{6.10}$$

where Q is the charge and E the electrical potential. This may be more practically written in terms of *power* since power is the measure of work expended in a unit of time:

$$\text{Power} = -EI \tag{6.11}$$

where E is electrical potential in volts and I is charge in amperes or charge per unit time, ${}^{\Delta Q}\!/_{\Delta t}$. Electrical power is often expressed in watts and 1 W-s is a joule. The sign is negative because the work is done by the system on the surroundings. To move one ampere of charge through a potential field of 1 V over 1 s will result in a work function of 1 J/s or 1 W. The work done when charge is separated and transferred is of great significance throughout biochemical processes.

6.4.4.2 Pressure–Volume Work

Most students are familiar with the expression of pressure–volume work, because the typically the teaching of thermodynamics is done with ideal gases and the concept of expansion work. Pressure–volume work can be important in biochemical processes especially in relation to water. Water expands when cooled to freezing, and work is done on the surroundings. Consider that the system under study is enclosed in a container with a wall that is movable, a piston. When this piston is moved up (expansion) or down (compression), work is done by or on the system with respect to the surroundings. There is a force per unit area, F/A, that is exerted on the system through the piston wall. This external force is the external pressure, P_{ext}. If the piston is moved a specific distance, dx, in the direction of the external force, a certain amount of work is done on the system:

$$dw = P_{\text{ext}}A\,dx \tag{6.12}$$

where A is the cross-sectional area of the piston. By integrating

$$w = \int P_{\text{ext}}A\,dx \tag{6.13}$$

There is a volume change in the system, dV that is described by the area of the piston moving through the distance dx:

$$dV = A\,dx \tag{6.14}$$

Substituting gives

$$w = -\int P_{\text{ext}}dV \tag{6.15}$$

The negative sign is explained as follows: The volume change will be negative for the compression, which gives a negative value to the work done by the compression. But, by convention, work done on the system must be positive, and so a negative sign must be added to the formula. This equation provides the work done on a system (when the pressure applied and the change in volume in the system are known). In SI units, the work done by PV work is in joules.

6.4.4.3 Gravity and Springs

We consider two major forms of mechanical work commonly encountered in biological systems. The first is the work done by moving a weight in a gravitational field. The second is work done when a force causes displacement of a structure acting like a spring. Consider an arm lifting a weight in the air. The action of the muscle fibers depends on work of the second kind, while the movement of the weight upward against gravity is a displacement of the first sort.

Gravitational work occurs when an object of mass m is lowered in a gravitational field (like that of the Earth) from a height h_1 to h_2. If the mass is lowered at a constant velocity, the following can be written as

$$w = mg\Delta h = mg\,(h_2 - h_1) \tag{6.16}$$

where g is the acceleration of gravity (9.8 m/s^2). Gravitational work is expressed in joules.

Processes in which the displacement occurs against a restoring force such as a spring obey *Hooke's law*. Hooke's law dictates that the force applied is directly proportional to the changed length of a spring. Consider the displacement of a spring whose length when no force is applied is x_o; when a force either compresses or expands the spring, the new length is x. Any force externally applied will be balanced by an internal force, the spring force. Hence, Hooke's law gives the following:

$$\text{Spring force} = -k\,(x - x_0) \tag{6.17}$$

where k is a constant for a given spring. The external force is equal and opposite to the spring force and is given by

$$\text{External force} = k\,(x - x_0) \tag{6.18}$$

Because the force necessary to stretch or compress a spring varies with the length of the spring for a change in length from x_1 to x_2, the work is calculated by integrating:

$$w = \int_{x_1}^{x_2} k\,(x - x_0)dx \tag{6.19}$$

which gives the result:

$$w = k\,(x_2 - x_1)\left(\frac{x_2 + x_1}{2} - x_0\right) \tag{6.20}$$

Hooke's law in many cases is a reasonable approximation not only for the behavior of interactions of muscle fibers and locomotive proteins in microtubules but also for the interactions of atoms with one another. It is a frequent and useful abstraction to consider atoms as connected by a spring.

6.4.4.4 Oscillatory Motion

So far we have treated only the restoring force and work done when a spring is compressed or stretched once. Springs have periodic motion when stretched and then allowed to come to rest. This oscillatory motion is characteristic of both springs and pendulums that are vibrating near their resting point. Pendulums, like springs, experience a restoring force given by **F** = –*kx*. The motion of an object (a ball) on a string or a spring can be derived from Newton's second law as follows:

First we determine the force acting on the ball which is the restoring force:

$$\mathbf{F} = ma = -kx \tag{6.21}$$

Because acceleration is the second derivative of position with respect to time (6.21) can be written as

$$\mathbf{F} = ma = m\frac{\partial^2 x}{\partial t^2} = -kx \tag{6.22}$$

In words, the second derivative of the ball's position is proportional to the position itself. (In this case the coefficient has a minus sign.) There are a limited number of functions that have this property. They are

$$\frac{\partial^2 \cos(\omega t)}{\partial t^2} = -\omega^2 \cos(\omega t) \tag{6.23}$$

$$\frac{\partial^2 \sin(\omega t)}{\partial t^2} = -\omega^2 \sin(\omega t) \tag{6.24}$$

There is a special name for functions with this property, namely that the derivative of a function returns the same function times a constant. Functions in this class are called *eigenfunctions*, *eigen* meaning "characteristic", "specific", or "innate." The general form of this solution is

$$x = A\cos(\omega t) = B\sin(\omega t) \tag{6.25}$$

where *A* and *B* are constants. The velocity and acceleration for the center of mass of a ball on a spring or a pendulum can be found by differentiating this equation:

$$v = \frac{dx}{dt} = -\omega A\sin(\omega t) + \omega B\cos(\omega t) \tag{6.26}$$

$$\begin{aligned} a &= \frac{\partial^2 x}{\partial t^2} = -\omega^2 A\sin(\omega t) + \omega^2 B\cos(\omega t) \\ &= \omega^2 x \end{aligned} \tag{6.27}$$

Combining Eqs. (6.22) and (6.27) gives the rate of oscillation in radians:

$$\omega = \sqrt{\frac{k}{m}} \tag{6.28}$$

A set of initial conditions can be chosen which then allow the constants *A* and *B* to be calculated: $x(0) = A$ and $v(0) = -\omega B$. The kinetic and potential energies can be calculated and shown to be conserved:

$$E_{\text{total}} = K + U = \frac{1}{2}kL^2 \tag{6.29}$$

L is the length of displacement. Figure 6.1 shows the motion of a system in harmonic motion along with the instantaneous changes in potential and kinetic energy. As the figure shows, the total energy remains constant as the potential and kinetic energies are exchanged. The highest velocity occurs when the mass moves through its point of zero displacement. If we were to take a picture randomly while the spring or pendulum is in motion, we would seldom see it at the point of zero displacement; instead we would find it where the kinetic energy is 0 and the potential energy is maximum: at the points of maximum displacement. As we will see, this is a classical result and is a special case, not the more general case expected by the rules of quantum mechanics (Chapter 8).

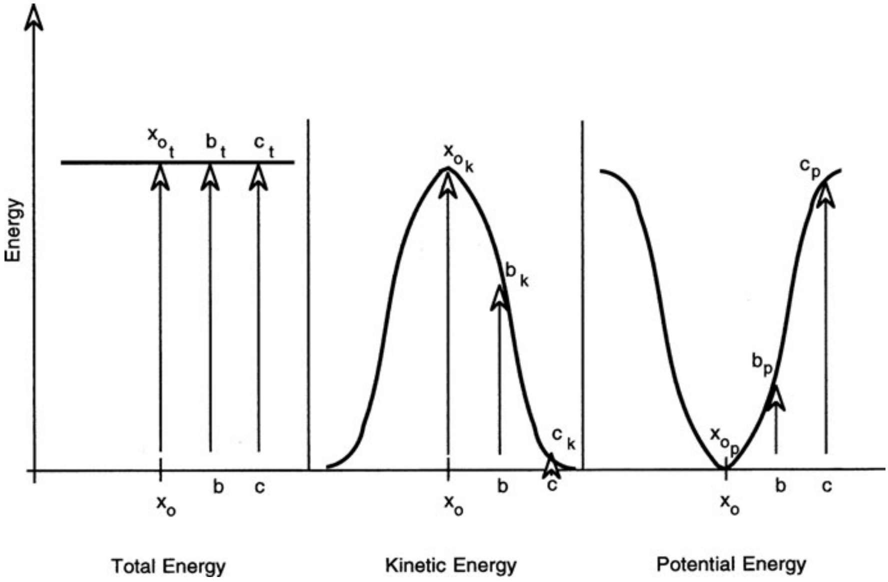


Fig. 6.1 Energy distribution at various points in a harmonic oscillator

Further Reading

Traditional physics texts can be helpful to extend the topics in this chapter and the next. The following are useful:

Feynman R.P., Leighton R.B., and Sands M. (1963) *The Feynman Lectures on Physics*, Volume 1–3. Addison-Wesley, Reading, MA. (This is a transcription of the lectures given by Feynman for a 2-year introductory physics course of his creation in the early 1960s at The California Institute of Technology. The style and insight almost always lead one to consult this classic text to find the deeper meaning of a physical principle.)

Halliday D., Resnick R., and Walker J. (2007) *Fundamentals of Physics*, 8th edition. Wiley, New York.

Tipler P.A. and Mosca G. (2007) *Physics for Scientists and Engineers*, 6th edition, W.H. Freeman, New York.

Warren W.S. (2000) *The Physical Basis of Chemistry*, 2nd Edition. Academic, San Diego. (This small book selects those topics in physics needed to understand most chemical processes. A quick easy read and well worth the time.)

The following texts are devoted to the application of physical principles in biological systems:

Hoppe W., Lohmann W., Markl H., and Ziegler H. (eds.) (1983) *Biophysics*. Springer-Verlag, New York.

Jackson M.B. (2006) *Molecular and Cellular Biophysics*. Cambridge University Press, Cambridge.

Nelson P. (2008) *Biological Physics: Energy, Information, Life*. W.H. Freeman, New York.

Phillips R., Kondev J., Theriot J., and Orme N. (2008) *Physical Biology of the Cell*. Garland Science, New York.

Sneppen K. and Zocchi G. (2005) *Physics in Molecular Biology.* Cambridge University Press, Cambridge.

Waigh T.A. (2007) *Applied Biophysics: A Molecular Approach for Physical Scientists.* Wiley, Chichester.

Weiss T.F. (1995) *Cellular Biophysics*, Volumes 1 and 2. MIT, Boston.

Problem Sets

1. Calculate the work performed by a battery delivering 100 mA at 9 V for 2 h. Express the answer in (a) joules, (b) calories, and (c) watt-hours.
2. What is the work performed by a 50 kg gymnast who performs a lift on the rings and is elevated 1.2 m in the air?

   (b) How many calories must be supplied to the muscles?
   (c) Assuming 100% efficiency in energy extraction from sugar and given that there are 4 kcal/g of sugar (5 g in a teaspoon, how many teaspoons of sugar should be consumed at lunch to make the lift?

Chapter 7
Biophysical Forces in Molecular Systems

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_7, © Springer Science+Business Media, LLC 2010

7.1 Form and Function in Biomolecular Systems Are Governed by a Limited Number of Forces

Position. . .Movement . . .linked through conservation of energy. If we conceive about the potential energy surface of a system as the set of forces that are responsible for *form* in a biological system, it is the movement or momentum over the surface, the kinetic energy that is *function*. Movement is kinetic energy; and motion is tied to spatial dimension as a function of time. The arrow of time is related to entropy and is the natural direction of all systems: time and energy, position and momentum. These fundamental properties of natural systems are the building blocks of biological form and function. We will see that these properties of systems are complementary and that they have very important philosophical and epistemological consequences in our understanding of biological systems. We cannot understand the biological system if we cannot appreciate its motion in state space. The central knowledge of biophysical chemistry is knowing which force induces motion over time and how such interactions work.

7.2 Mechanical Motions Can Describe the Behavior of Gases and the Migration of Cells

The study of motion is the study of mechanics. Whether the motion is an actin molecule or an atomic orbital, there are just several basic motions that sum up the majority of the mechanical actions in biological systems. These are

1) linear motion in constant potential space.
2) linear motion in a varying potential (acceleration).
3) circular motion in a constant field (the centrifuge – Chapter 28).
4) harmonic motion in a parabolic field (Hooke's law – Chapter 6).
5) motion in fields of varying potential (boxes, wells, spherical wells, non-linear potentials).

Knowledge of these types of motion will be used over and over in our exploration of molecular biophysics. As we progress through this book we will selectively review aspects of classical mechanics that are important for understanding that topic as needed. Here we will review the consideration of molecular motion important in the treatment of gases and ideal materials.

7.2.1 Motion in One and Two Dimensions

Generally in mechanics we are not concerned with the inner structure or workings of an object in motion and so the most common *abstraction* introduced is that of a *particle*. A particle is often considered as something very small, but size as such is relative. At the quantum level a nucleus with the size of 10^{-15} m is not considered a particle if we are interested in its internal makeup and behavior. Conversely we may well consider the Milky Way Galaxy as a particle if the observables of the system under consideration are not affected by the intrinsic organization of the Milky Way. As we will see, complex objects can often be treated as a system of particles and this extension of the particle abstraction is an extremely valuable one.

Initially we describe motion of a single particle in one dimension (Table 7.1). The *motion* of a particle is described in terms of the *position* of a particle which requires it to be located in an appropriate coordinate system. A particle that is in motion undergoes *displacement* within this coordinate system. The rate of this displacement with respect to time is the *velocity*. The velocity can also change from moment to moment and a change in the rate of the velocity of a particle is called its *acceleration*. If these quantities are expressed as an average over a period of time we write them with a Δ sign; but if the full information of the instantaneous displacement, velocity, and acceleration are necessary to the description of our sample space, we must use the mathematical techniques of calculus. It was for the purpose of expressing the instantaneous motion of particles that Newton invented calculus. We can define the motion of a particle in one dimension only along a single line. The use of a positive or negative sign is adequate to identify direction. In two or more

Table 7.1 Elementary motions in one dimension

	Average value	Instantaneous value
Displacement	$\Delta x = x_2 - x_1$	dx
Velocity	$v_{\text{ave}} = \dfrac{\Delta x}{\Delta t}$	$v(t) = \dfrac{dx}{dt}$
Acceleration	$a_{\text{ave}} = \dfrac{\Delta v}{\Delta t}$	$a = \dfrac{dv}{dt} = \dfrac{d(dv/dt)}{dt} = \dfrac{d^2x}{dt^2}$

The instantaneous displacement is related to the time of displacement. The velocity is the first derivative of displacement with respect to time and acceleration is the second derivative of displacement with respect to time. The average velocity is the area under the curve found by integrating $\frac{dx}{dt}$ in the interval from t_2 to t_1

dimensions we use vectors to define the movement of particles since they possess both magnitude and direction.

When motion in more than a single dimension is considered, we consider displacement, velocity, and acceleration as *vectors* and use the appropriate vector mathematics to describe their interactions. An important but unfortunately confusing aspect of working with vectors is the definition of position in space that in three dimensions requires three coordinates per particle. The choice of coordinate system is usually a matter of convenience. For example, when a system contains spherical symmetry, it is more convenient to use a spherical coordinate system than one that is rectangular. (The convenience can be appreciated in the case of a particle moving on the surface of a sphere in which case the coordinate r will be constant for the problem with only θ and ϕ varying.) In general, problems in multiple dimensions are considered by separating a motion into its component vectors that are directed in one-dimensional fashion. These vectorial components are then related to one another and to the net resultant by the techniques of trigonometry and vector mathematics. Multiple examples of this analysis can be found in standard undergraduate textbooks of physics and will not be explored in detail here except to give the mathematical results of several of the more common cases that are applicable to our interests.

7.2.2 Motion Under Constant Acceleration

The motion of a particle undergoing constant acceleration, a_{o}, is a common feature in many natural systems. The free fall of a particle in a gravitational field is perhaps the most common example where the acceleration will be equal to ($g = 9.8\ \text{m/s}^2$). If no retarding forces are considered (air resistance, for example) and no other forces are present the equations of displacement and velocity are written:

$$v(t) = v_{\text{o}} + a_{\text{o}}t \tag{7.1}$$

$$x(t) = x_{\text{o}} + v_{\text{o}}t + \frac{1}{2}a_{\text{o}}t^2 \tag{7.2}$$

Since time and displacement are related, we can write t in terms of x and eliminate it from these equations:

$$v^2 = v_o^2 + 2a_o(x - x_o) \tag{7.3}$$

In all of these equations the subscript "o" indicates the initial conditions, v is the velocity, t is the time, and x, y, z are coordinates in a Cartesian coordinate system.

7.2.3 Projectile Motion in a Constant Potential Energy Field

The classical example of this motion is an object fired horizontally from a mountain near the Earth surface. The field is characterized by vertical acceleration, g. This is the equation we show here because it is familiar from previous studies in physics. We will be more concerned with the motion of a charged particle such as an ion moving in a field of constant electrical or magnetic potential. In those cases, which we will see in later chapters, the constant acceleration caused by the constant field will exchange the term g with the more general one, a.

The components of the velocities of a projectile fired at a constant velocity v_x, perpendicular to the constant potential field (the minus sign says the particle is attracted toward the origin of the field) will be found by:

$$v_x = v_{o(x)} \text{ and } v_y = v_{o(y)} - gt \tag{7.4}$$

This allows the computation of position at any time. In terms of the components

$$\begin{aligned} x(t) &= x_o + v_{o(x)}t \\ y(t) &= y_o + v_{o(y)}t - \tfrac{1}{2}gt^2 \end{aligned} \tag{7.5}$$

The ability to describe the motion of a system of particles allows the computation of the mechanical actions that characterize the properties of systems of chemical interest. We can develop this ability by starting with the simplest case, the actions that lead to a description of the properties of an ideal gas.

7.3 The Kinetic Theory of Gases Explains the Properties of Gases Based on Mechanical Interactions of Molecules

In the classical ideal system, defined as one in which individual particles do not interact with each other, all we need to know is the initial position and momentum of each particle. Each particle's behavior is given by the mechanics that we have just described; the possible arrangement of initial positions and momenta is given by a distribution function (see Chapter 5); and the behavior of the system is simply the sum of the interactions. The simplest example of this type of analysis is the kinetic theory of gases.

7.3.1 Collisions Are Impacts in Which Objects Exchange Momentum

We have discussed the conservation of total energy in a system as the kinetic and potential energy are exchanged as a consequence of position and motion. The *collision* of particles is a very important action in chemical studies and is best understood in terms of the exchange of momentum (conservation of momentum). If two particles are moving toward one another each with a certain momentum, and meet, a collision takes place. A collision is an impact in a brief interval in which otherwise independent objects exert strong forces on one another. The force associated with a collision is called an *impulse*. Whereas work depends on the integral of a force over distance (change in coordinates in space), an impulse depends on the integral of the force over a time interval. If two objects collide the total kinetic energy may or may not be conserved. When KE is conserved the collision is said to be *elastic*. Examples of elastic collisions include the exchange of velocities when two bodies of equal mass collide such as in billiards: the traveling cue ball will stop and the eight ball will move in the same trajectory with the same velocity (assuming one-dimensional travel). Two balls moving toward each other will recoil back the way they came from with the other ball's speed. If a particle of small mass strikes a much more massive object, the smaller particle will reverse direction leaving with almost the same speed, and the larger object will move very slowly in the opposite direction. This is like the case of the billiard ball striking the bumper and like the case of gas molecules inside a container (which we will explore shortly). In an elastic collision the trajectory of the center of mass for the system is not altered by the collision. The general conservation equations are

for momentum

$$m_1v_1 + m_2v_2 = m_1v_1' + m_2v_2' \tag{7.6}$$

and for kinetic energy

$$\begin{aligned} K.E. &= \frac{m_1\,|v_1|^2 + m_2\,|v_2|^2}{2} \\ &= \frac{m_1\,\left|v_1'\right|^2 + m_2\,\left|v_2'\right|^2}{2} \end{aligned} \tag{7.7}$$

Collisions in which the total energy remains the same but the kinetic energy is not conserved are called *inelastic*. A totally inelastic collision occurs when two particles stick together. Inelastic collisions can occur with an increase in kinetic energy (an explosion) or with a decrease in kinetic energy (usually an increase in the internal energy of the particles).

7.3.2 Reviewing the Phenomenology of Dilute Gases Sheds Light on Molecular Mechanics

Among the earliest quantitative studies of molecular systems were experiments on the pressure, volume, and temperature relationships in dilute gases. In 1662, Robert Boyle explored the relationship between the volume, V, and pressure, P, of a fixed amount of gas at a constant temperature. Boyle discovered that under these constraints of fixed mass and temperature, the volume and pressure were inversely proportional

$$V \propto \frac{1}{P} \tag{7.8}$$

which is equivalent to

$$VP = \text{constant} \tag{7.9}$$

This expression is *Boyle's law*. Boyle's law gives rise to the important relationship that for a fixed quantity of gas, n, at a constant temperature or isothermal condition:

$$P_1V_1 = P_2V_2 (n, T \text{ constant}) \tag{7.10}$$

If all of the various combinations given by Eq. (7.10) are plotted, Fig. 7.1 is the result. The line connecting all of the PV terms whose product is constant is called an *isotherm*. The form of the isotherm for Boyle's law is a hyperbola.

What happens if the temperature of the gas is varied? Experiments to investigate what happens to the volume of a gas when the temperature is varied at constant pressure and mass were done by Alexander Charles (1746–1823) and Joseph Louis Guy-Lussac (1778–1850) in the early nineteenth century. (Their interest was to improve the performance of hot air balloons, which were popular at that time.) These studies showed that when pressure and mass of a gas were constrained, the volume and temperature varied proportionately:

$$\begin{gathered} V \propto T \\ \frac{V}{T} = \text{constant} \end{gathered} \tag{7.11}$$

This is the law of Charles or Charles and Guy-Lussac. A gas, when heated expands and when cooled contracts, to occupy a larger or smaller volume, respectively. Experiments show a linear relationship of the change in volume with respect to temperature when the pressure is held constant as shown in Fig. 7.2.

How can temperature be quantified? An important operational definition of temperature can be made beyond a measure of “hotness” or “coldness” using these properties of gases. Imagine that two identical containers, A and B, each with flexible walls that can expand or contract so that the volume varies such that the pressure exerted on the walls remains constant, are filled with an identical fixed amount of

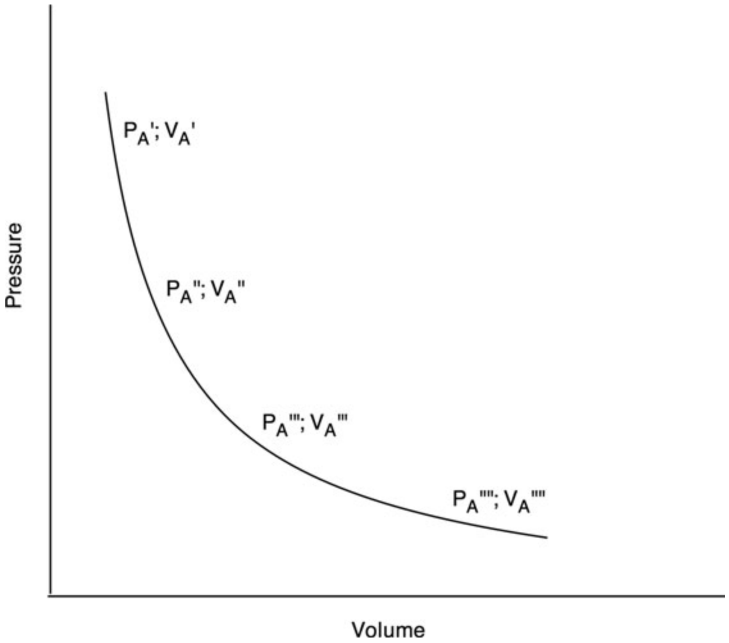


Fig. 7.1 Boyle’s law generates a plot of pressure versus volume at constant temperature and constant amount of gas. This plot is an isotherm

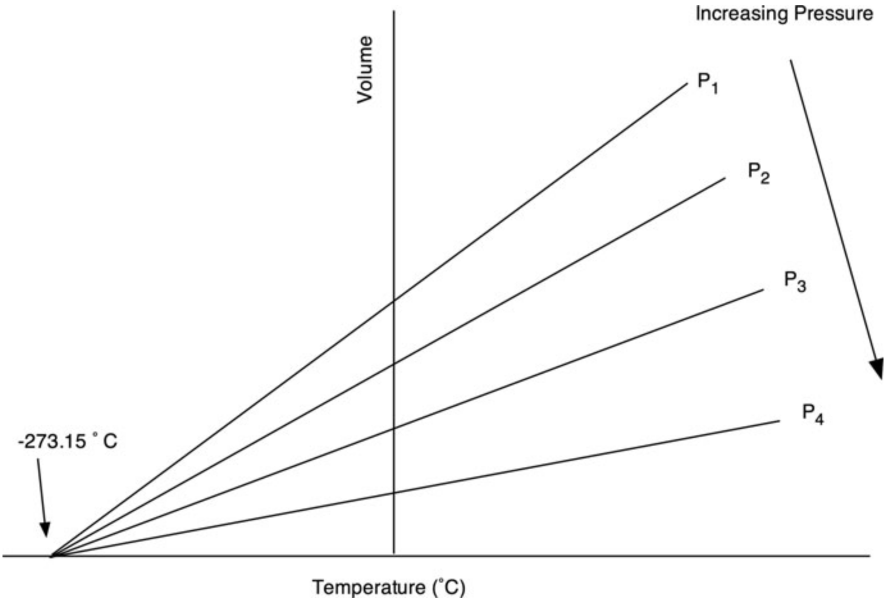


Fig. 7.2 Charles law and the determination of absolute zero can be seen when the volume of a gas is plotted against a changing in temperature. The amount of the gas is fixed and the relationship is determined at various pressures. All of the generated straight lines intersect with the *y*-axis at a common point which represents zero volume and is at −273.15°C

gas. Container A is characterized by $V_A P_A$ and container B by $V_B P_B$. Their initial temperatures may vary. The containers are now brought into contact so that they can exchange heat across their walls (but nothing else) and are allowed to come into thermal equilibrium (the same temperature): A is now at the state, $V'_A P'_A$, and B is at the state, $V'_B P'_B$. Following the Boyle's Law relationship, we could now uncouple A from B, change the VP state of A to $V''_A P''_A$ or $V'''_A P'''_A$ or any of an infinite number of other $V^x_A P^x_A$and still have A in thermal equilibrium with B. All of the possible various VP states of A that will fulfill this thermal equilibrium condition are described by the isotherm in Fig. 7.1. Since all of these various states of A are in thermal equilibrium with the state of B, $V'_B P'_B$, they share a variable in common that is temperature, T. This experiment can be generalized to the following phrase, known as the *zeroth law of thermodynamics*:

> when two systems are in thermal equilibrium with a third system, then they must be in thermal equilibrium with each other.

The zeroth law of thermodynamics is the basis on which thermometry is founded. Each different temperature will form a unique isotherm in a PV graph as drawn in Fig. 7.1.

The concept of an absolute temperature scale and the assignment of value to absolute zero can be derived from these three relationships. As Fig. 7.2 shows, each of the unique straight lines relating volume-to-temperature that are found at various pressures can be extrapolated. All of the lines meet at the x intercept. On a thermal scale of degrees centigrade this x-intercept will be valued at –273.15°C. In actual experiments, all gases will become liquid before this temperature is reached so the line must be extrapolated to actually reach the theoretical point of zero volume given by the law. The implications of this zero volume temperature were pointed out by William Thomson (Lord Kelvin) in 1848. He noted that this temperature is the theoretically lowest attainable temperature and created a temperature scale with this absolute zero as the starting point. The absolute temperature scale is called the Kelvin temperature scale and one Kelvin (K) is equal in magnitude to 1° centigrade. When temperature appears in equations in this volume and in all thermodynamic equations it must always be in K for the correct magnitudes associated with the gas and Boltzmann constants to be properly used.

A remarkable discovery connecting the amount of a gas with its volume, temperature, and pressure was made by Amedeo Avogadro in 1811. Avogadro's law (Eq. (7.10)) says that regardless of identity, equal volumes of any gas at the same temperature and pressure contain the same number of molecules. This is written

$$V \propto n \text{ or } \frac{V}{n} = \text{constant} \tag{7.12}$$

At the standardized conditions of 1 atm of pressure and 273.15 K (standard temperature and pressure or STP), 1 mol of gas occupies 22.414 l.

Boyle's, Charles', and Avogadro's laws are all interrelated and can be combined into an equation of state just as we have discussed in Chapter 4. Each of the laws has been written above in terms of volume. Combining these yields

$$V \propto \frac{nT}{P} \tag{7.13}$$

If a proportionality constant can be found, this expression can be written as an equality:

$$\begin{aligned} V &= R\frac{nT}{P} \\ PV &= nRT \end{aligned} \tag{7.14}$$

This is the familiar *ideal gas equation*. The values of *R*, the gas constant, can be calculated from the experimental values. The ideal gas equation, though drawn from experimental data, makes several significant assumptions. These are the assumptions of ideality: (1) the particles making up the system do not occupy significant space and (2) there are no interactions between particles of an attractive or repulsive nature (i.e., collisions, if they were to occur would be perfectly elastic). We can now explore the kinetic theory of gases with respect to the results of the ideal gas equation. Following that we will examine more complete models for systems of molecules that account for the real deviations from ideality that we often discover.

7.3.3 The Pressure of a Gas Is Derived from the Transfer of an Extremely Small Amount of Momentum from an Atom to the Wall of a Vessel

We can deduce the observable nature of a gas in terms of its pressure and temperature simply by considering a large number of small particles (atoms) interacting with a very massive object (the walls of a container) and applying Newton's laws of motion, principally conservation of momentum. In other words we consider the collisions of a large number of atoms within a constrained volume. This analysis leads to the kinetic theory of gases which when coupled with the Boltzmann distribution allows us to obtain the macroscopic observables of pressure and temperature of a gas. Such an analysis also provides a link between classical and statistical thermodynamics.

We start by considering a helium atom (mass $= 6.6 \times 10^{-27}$ kg) moving at a speed of 1200 m/s perpendicular to the wall of a container that weighs 1 kg. Its momentum is $\mathbf{p} = 7.92 \times 10^{-24}$ m kg/s. After striking the wall, elastically it will be moving in the opposite direction with $\mathbf{p}' = -7.92 \times 10^{-24}$m kg s^{-1}. How much momentum is transferred to the vessel wall upon collision? We use Eqs. (7.6) and (7.7) and note that the “′” represents a value following collision with the wall:

$$\begin{gathered}
\mathbf{p}_{\mathrm{He}} + \mathbf{p}_{\mathrm{wall}} = \mathbf{p}'_{\mathrm{He}} + \mathbf{p}'_{\mathrm{wall}} \\
\left(7.92 \times 10^{-24}\mathrm{m\ kg\ s^{-1}}\right) + (0) = \left(-7.92 \times 10^{-24}\ \mathrm{m\ kg\ s^{-1}}\right) + \left(\mathbf{p}'_{\mathrm{wall}}\right) \\
\left(1.584 \times 10^{-23}\ \mathrm{m\ kg\ s^{-1}}\right) = \mathbf{p}'_{\mathrm{wall}} = \left[(1\ \mathrm{kg})\left(v'_{\mathrm{wall}}\right)\right] \\
v'_{\mathrm{wall}} = 1.584 \times 10^{-23}\ \mathrm{m\ s^{-1}}
\end{gathered} \tag{7.15}$$

This equation shows that for an elastic collision with the wall, the momentum exchange leaves the wall with a very small velocity. The velocities of the helium atom and the wall allow the calculation of the kinetic energy possessed by each:

$$E_{\mathrm{He}} = \frac{m_{\mathrm{He}} v_{\mathrm{He}}^2}{2} = \frac{\left(6.6 \times 10^{-27}\mathrm{kg}\right)\left(1200\mathrm{m\ s^{-1}}\right)^2}{2} = 4.752 \times 10^{-21}\mathrm{J} \tag{7.16}$$

$$E'_{\mathrm{wall}} = \frac{m v'^2_{\mathrm{wall}}}{2} = \frac{(1\ \mathrm{kg})\left(1.584 \times 10^{-23}\ \mathrm{m\ s^{-1}}\right)^2}{2} = 1.254 \times 10^{-46} J \tag{7.17}$$

Only by leaving most of the kinetic energy in the atom can both conservation laws be accorded. Thus the atom bounces off the wall with an elastic collision with virtually the same magnitude velocity that it arrived, 1200 m/s. Though virtually all of the energy remains in the atom, a very small amount is imparted to the wall, approximately one part in 2.6×10^{-26}. This small energy transfer is the source of the pressure exerted by the gas on the vessel wall.

We can now directly determine the pressure of a gas in a cubic volume (length $= l$) using the mechanics of elastic collisions. Pressure is defined as the force per unit area; and the momentum imparted to the vessel wall derived from the collisions of molecules in a gas causes this force. We consider only an ideal gas at low pressure that has the properties:

- it is mostly empty space;
- The molecules are non-interacting (i.e., the intermolecular forces are negligible and collisions between molecules are also neglected);
- the molecules are in constant motion with a distribution of speeds and directions that are random and depend on the kinetic energy of the group.

As our previous example showed, when a single gas molecule moving in a direction perpendicular to a side of the box hits the wall, there is a momentum change imparted to the wall of $\Delta\mathbf{p} = 2mv$. As we saw, the velocity of the molecule in the limit is unchanged, and every collision with the wall requires a round trip (e.g., twice the distance l) that takes time: $\Delta t = 2l/v$. The average force on one wall will be the momentum transferred per unit time:

$$\mathbf{F} = \frac{\Delta \mathbf{p}}{\Delta t} = \frac{2mv}{2l/v} = \frac{mv^2}{l} \tag{7.18}$$

The pressure, P, is the force per unit area

$$P = \frac{\mathbf{F}}{A} = \frac{mv^2/l}{l^2} = \frac{mv^2}{l^3} = \frac{mv^2}{V} \tag{7.19}$$

because $l^3 = \text{Volume}$. For a container of 1-l volume (10 cm^3) with the helium atom of the previous example the pressure of a single atom would be very small (on the order of 10^{-20}). If there are N molecules, they will exert approximately N times Eq. (7.19):

$$P = N\frac{m}{V}\left\langle v^2 \right\rangle \tag{7.20}$$

We say "approximately" because not all of the molecules in N will strike on the y walls; some will be striking the x and z walls as well. This will not change the rate of collisions nor the momentum imparted to the y-axis side. Thus as long as the density of the gas is kept constant, Eq. (7.20) is independent of the initial direction of the velocity vector as well as the size and shape of the box. The velocity distributions in each x, y, and z direction are equal (actually 1/6th of the atoms strike each wall, 1/3 for each coordinate)

$$\left\langle v_x^2 \right\rangle = \left\langle v_y^2 \right\rangle = \left\langle v_z^2 \right\rangle \tag{7.21}$$

$$\left\langle |\mathbf{v}|^2 \right\rangle = \left\langle v_x^2 + v_y^2 + v_z^2 \right\rangle = 3\left\langle v_y^2 \right\rangle \tag{7.22}$$

The total energy is equal to the sum of all of the particles:

$$E = \sum_{i=1}^{N} \frac{m\,|\mathbf{v}|^2}{2} \tag{7.23}$$

or

$$PV = \frac{2}{3}E \tag{7.24}$$

Finally, since the molecules do not all have the same speed we should take the average speed of all of the molecules. The distribution of the molecules' speed is derived from the Boltzmann distribution and is the Maxwell–Boltzmann distribution function. This is a Gaussian function for the number of particles found at each speed depends on temperature (Fig. 7.3). For a single dimension the average velocity is

$$\left\langle v_y^2 \right\rangle = \frac{kT}{m} \tag{7.25}$$

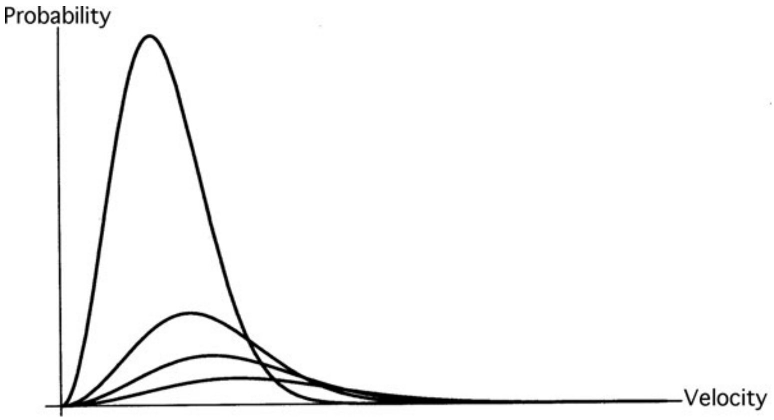


Fig. 7.3 Maxwell–Boltzmann distributions for the molecular speed distributions for oxygen at 100, 200, 273, and 373 K (*left* to *right*)

k is the Boltzmann constant and *T* is the absolute temperature. Since just as many particles are going left as right, the root mean square of the velocities is used to keep the average from coming out to zero:

$$< \mathbf{v}_{\mathrm{y}}^{2} >^{1/2} = \sqrt{\frac{kT}{m}} \tag{7.26}$$

In three dimensions the velocity distribution is

$$< \mathbf{v}_{\mathrm{y}}^{2} >^{1/2} = \sqrt{\frac{3kT}{m}} \tag{7.27}$$

Combining Eqs. (7.23) and (7.27) yields

$$E = \frac{3}{2}NkT \tag{7.28}$$

This equation gives us an important perspective on temperature. The temperature of a state is related to both its internal energy and the motion of its atoms. The classical result is that a temperature of absolute zero (0 K) means that the kinetic energy is 0 and the molecules are completely motionless.

Applying the results from the treatment of the gas above, we can see that

$$PV = \frac{2}{3}E_k = nRT \text{ or } E = \frac{3}{2}nRT \tag{7.29}$$

with $E_k = N\left(\frac{1}{2}mv^2\right)_{\text{average}}$ which is total translational kinetic energy of the molecules. If this translational energy is considered the total internal energy of the gas (i.e., no forms of energy other than the kinetic energy of translation are considered), then the internal energy U will depend only on temperature and not on pressure and volume. By restricting all other forms of energy in the system, we also restrict all interactions between the particles in the system since only elastic collisions are allowed. This is the definition of an ideal gas (or, as we will see later any ideal phase).

$$PV = nRTU = U\,(T) \text{ only, and } U = \frac{3}{2}nRT \tag{7.30}$$

with n and R defined as usual. Once the internal energy and the constraint of temperature are fixed, then all other state properties, in this case PV is also fixed.

7.3.4 The Law of Equipartition of Energy Is a Classical Treatment of Energy Distribution

The law of equipartition of energy or the *equipartition theorem* results from a classical statistical–mechanical argument based on the Boltzmann distribution and states

> *For a system at thermal equilibrium, T, an average energy of $\frac{kT}{2}$ is assigned to each degree of freedom per molecule, where k is Boltzmann's constant.*

The equipartition theorem results from the kinetic treatment of a classical system of gas molecules but can be applied to any classical system containing a large number of identical entities. Practically speaking an equation for the total energy of a system is written and a value of $\frac{kT}{2}$ is assigned to each squared term of the energy equation. From a physical standpoint this information can be used to infer a structure from measurements related to the internal energy of the entities comprising the system (this is the essence of much of thermodynamics and will be reviewed in some detail in Chapters 10–13).

Problem: Use the equipartition theorem to argue that nitrogen and oxygen exist as diatomic gases.

Answer: Consider the diatomic molecule as a rigid-dumbbell model with translational capacity in three dimensions and with the ability to rotate in the x and y planes (Fig. 7.4). Rotation around the z axis is restricted because the rigid dumbbell lies in the z axis and prevents any change in energy following a collision.

The kinetic energy can thus be written as

$$E_k = \frac{1}{2}mv_x^2 + \frac{1}{2}mv_y^2 + \frac{1}{2}mv_z^2 + \frac{1}{2}I_{x'}\omega_{x'}^2 + \frac{1}{2}I_{y'}\omega_{y'}^2 \tag{7.31}$$

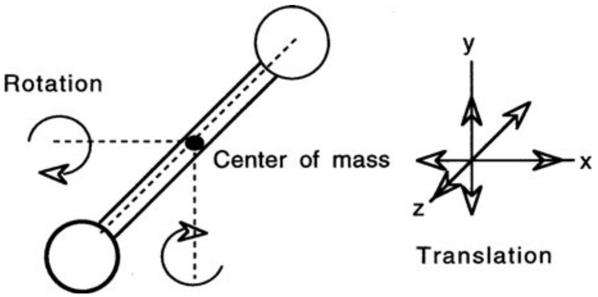


Fig. 7.4 The degrees of freedom in which a contribution to the kinetic energy of the molecule is made: There are five in this model of a rigid-dumbbell model

There are five degrees of freedom in this system each identifiable by the squared term in the energy equation therefore:

$$\overline{E} = N\,(5)\,\frac{kT}{2} \tag{7.32}$$

This can be put on a molar basis by realizing that $N = nN_A$ where n is moles and N_A is Avogadro's number and that $N_A\ k = R$. Thus

$$\overline{E} = \frac{5}{2}nRT \tag{7.33}$$

We will see in Chapter 10 that the change in internal energy with respect to changes in temperature is called the *heat capacity*. This measurement can provide an empirical tool to test the proposed model. Therefore, for a system held at constant volume

$$C_v = \frac{dE}{dT} = \frac{5}{2}nR \tag{7.34}$$

At constant pressure, C_v is 20.785 J/mol K, which, when compared to the heat capacities listed in Table 7.2, is the value measured for the diatomic gases N_2 and O_2. The model proposed for a diatomic system is therefore supported by the measurements and linked by the formal model described by the equipartition theorem. This analysis is the basis for Clausius' postulate made in 1880 that oxygen and nitrogen are diatomic gases capable of rotation around two axes. We can use the equipartition theorem to argue that diatomic molecules do not vibrate.

> *Problem*: How would the equipartition equation change for diatomic molecules that vibrate? Comparing this equation to the data in Table 7.2, what does this predict about diatomic molecules?

Table 7.2 Heat capacities for selected materials. 25°C, constant volume

Material	State	C_{v}^{-} (J/K/mol)
He	Gas	12.5
H_2	Gas	20.5
O_2	Gas	21.1
N_2	Gas	20.5
H_2O	Gas	25.2
CH_4	Gas	27.0
CO_2	Gas	28.5
H_2O	Liquid	75.2
Ethanol	Liquid	111.4
Benzene	Liquid	136.1
Cyclohexane	Liquid	183.3
n-Hexane	Liquid	195.0

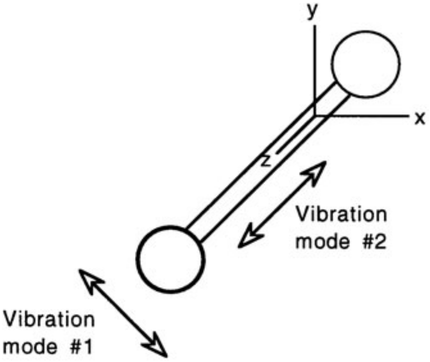


Fig. 7.5 Degrees of freedom available to the dumbbell model of a diatomic molecule

Solution:

As Fig. 7.5 shows, two more degrees of freedom are available to the diatomic molecule if vibrational modes are considered. Hence the expression for the energy would be

$$\overline{E} = \frac{7}{2}nRT \tag{7.35}$$

However, once again, the comparison with the measurement of heat capacities suggests that diatomic molecules do not vibrate.

The equipartition theorem is limited, and these limitations lie in its classical assumptions that any continuous value may be taken by the entities in a system. The equipartition theorem predicts that there should be no temperature dependence for heat capacities and provides no insight into why diatomic molecules do not vibrate. Looking ahead, the reason for the failure is that atomic systems are *not* permitted to take any continuous value for energy assignments. A new mechanical formalism, quantum mechanics, in which the possible values for energy assignments are quantized, provides the explanation. The energy jumps allowed for vibration turn out to be on the order of several tenths of an electron volt ($\approx 1.6 \times 10^{-19}$ J) which is significantly higher than the energy available through collisions, $kT = 4.28 \times 10^{-21}$ J, at room temperature. Consequently, these energy transitions are essentially forbidden at room temperature and because they simply are not seen, the vibrational modes of diatomic molecules do not play a role in the internal energies of diatomic molecules at low temperatures.

7.3.5 The Real Behavior of Gases Can Be Better Modeled by Accounting for Attractive and Repulsive Forces Between Molecules

The ideal gas equation clearly does not capture the complete behavior of any gas since it predicts that at 0 K the volume of a mol of gas would be 0 L. Not only is this not true, all gases change state to liquid before reaching 0 K. For the ideal gas law this is a bifurcation/surprise in the predicted state space that requires revision of the assumptions on which the equation of state is built. To restate these assumptions: (1) the gas molecules have negligible volume and (2) that there are no interactions, attractive or repulsive, between the molecules (a more complete list of the assumptions is given in Appendix E). The conditions under which these assumptions are realized are (1) in dilute gases (i.e., pressures <10 atm) where the relative volume of the molecule is much, much less than the total volume it occupies, and (2) at high temperatures where interaction energies overcome by the kinetic energy of the molecules. These factors lead to substantial deviation from ideal behavior in actual experimental behaviors.

When a gas is compressed or cools, the volume falls and the molecules of a gas interact leading to significant deviations from ideality. As molecules come close to one another, their actual size leads them to physically interact through both attractive and repulsive forces. Very close contact causes repulsive forces to dominate while attractive forces typically have effects at longer range. As we will see in coming chapters, the balance between these two classes of forces determines the behavior of biomolecular systems. Two important approaches to the modeling the behavior of real gases are the van der Waals equation and the virial equation of state.

7.3.5.1 The van der Waals Equation Is a Mathematical Model that Captures Non-ideal Behavior in a Gas

In 1873, Johannes Diderik van der Waals (1837–1923) modified the ideal gas equation by taking the approach that the ideal gas law could be a better model with the addition of specific factors. He made attempts to account for both the finite volume of the individual molecules and the attractive forces between them. The finite volume of the molecules would become most apparent in the cases where the ideal gas equation predicts a vanishing volume, thus a factor reflecting a *higher* volume than predicted at low temperatures and high pressures will be needed.

$$\begin{aligned}\text{volume occupied by molecules } &\propto n \\ &= nb\end{aligned} \tag{7.36}$$

b is the proportionality constant that gives nb units of volume, thus b has the dimensions of L/mol.

Furthermore the derivation of the pressure from a gas by the kinetic theory calculated the pressure from the frequency and number of molecules striking the walls of the container in a particular time. Attractive forces that would make the molecules stick to one another (i.e., inelastic collisions) will decrease the frequency of strikes and the average number of molecules available to make the wall-strikes. Therefore a factor is needed that modifies the apparent pressure by *reducing* it below the prediction of the ideal gas equation under conditions where inelastic collisions are more likely (again low temperature and high pressure). Both of these cases reduce the number of available molecules for wall-collision or the effective density of the gas, n/V. We can write

$$\begin{aligned}\text{pressure reduction } &\propto \left(\tfrac{n}{V}\right)\left(\tfrac{n}{V}\right) \\ &= a\left(\tfrac{n}{V}\right)^2\end{aligned} \tag{7.37}$$

where a is the proportionality constant for the attractive forces between the molecules and has the dimensions of atm L^2/mol.

These adjusted factors are combined with the ideal gas law to give the van der Waals equation:

$$\left(P + a\left(\frac{n^2}{V^2}\right)\right)(V - nb) = nRT \tag{7.38}$$

The constants a and b have positive values and are specific to each individual gas (Table 7.3). The equation as it is composed uses P and V as the measured pressure and volume for a system. The first term $\left(P + a\left(\frac{n^2}{V^2}\right)\right)$ is the pressure without attractive forces and the second term $(V - nb)$ is the volume without correction for the excluded volume of the molecules. Thus in the limit of a and b approaching zero, the van der Waals equation of state approaches the ideal gas law.

The van der Waals equation is a better model for the behavior of gases in a wider array of conditions. The validity of a as a measure of attractive forces between

Table 7.3 van der Waals coefficients

Gas	a (Pa m^3)	b (m^3/mol)
Helium	3.46×10^{-3}	23.71×10^{-6}
Neon	2.12×10^{-2}	17.10×10^{-6}
Hydrogen	2.45×10^{-2}	26.61×10^{-6}
Carbon dioxide	3.96×10^{-1}	42.69×10^{-6}
Water vapor	5.47×10^{-1}	30.52×10^{-6}

molecules is strengthened by the relationship between *a* and the boiling points of liquids. We would expect substances with increased intermolecular attractions to have higher boiling points (and larger values for *a*). This can be demonstrated. The excluded volume term, *b*, is also useful but does not always correlate to the size of molecules. Thus while the van der Waals equation provides both a molecular explanation and a better model of the real behavior of gases compared to the ideal gas law, further modeling work is required for a more complete understanding.

7.3.5.2 The Virial Equation Is Mathematically Powerful but Requires Empirical Fitting of Its Coefficients

van der Waals took a mechanistic approach to improving the ideal gas law's capacity for capturing real behavior. Another approach also used in biophysical studies is to use the *virial equation of state*. The term virial is derived from the latin and means "strength, force or energy." The virial equation is probably the most versatile and accurate mathematical equation of state and is derived from a power series expansion of the variable n/V or the molar volume. The deviation from ideal behavior of a gas can be experimentally expressed in terms of a compressibility factor, Z and is written as

$$\begin{aligned} Z &= \frac{PV}{nRT} \\ &= \frac{P\overline{V}}{RT} \text{ (in terms of molar volume)} \end{aligned} \tag{7.39}$$

For an ideal gas, Z is equal to one regardless of the pressure. For a real gas that has intermolecular interactions, Z will vary with pressure. The virial equation expands Z in terms of the inverse powers of $\overline{V}$:

$$\frac{P\overline{V}}{RT} = Z = 1 + \frac{B}{\overline{\overline{V}}} + \frac{C}{\overline{\overline{V^2}}} + \frac{D}{\overline{\overline{V^3}}} K \tag{7.40}$$

all of the coefficients are dependent on temperature. The first is one and B, C, D, . . . are the second, third, and fourth, etc., so-called virial coefficients. For an ideal gas all of the coefficients higher than one are equal to zero. The coefficients can be evaluated by *P–V–T* data of a gas with curve-fitting to any degree of accuracy simply by adding terms as needed. The second viral coefficients have been experimentally determined for some gases but not many substances have coefficients determined for the third and higher coefficients. The coefficients can be derived theoretically from intermolecular potential functions, however.

Alternatively, Z can be expanded in terms of pressure

$$Z = 1 + B'P + C'P^2 + D'P^3 + K \tag{7.41}$$

These equations (7.40 and 7.41) can be equated and the relationships the two sets of virial coefficients can be found

$$B' = \frac{B}{RT}$$

$$C' = \frac{C - B^2}{(RT)^2} \tag{7.42}$$

$$D' = \frac{D - 3BC + 2B^2}{(RT)^3}$$

The relationships between the coefficients B', C', D' usually allow (7.41) to be written in only the first two terms. The second virial coefficient, B, can be interpreted as representing the intermolecular mechanisms described in the van der Waals equation and they are related as follows:

$$B = b - \frac{a}{RT} \tag{7.43}$$

While the challenge of using the virial expansion is always to interpret the coefficients that are added in molecular terms, the approach is important for several reasons. First, it is flexible and can be made as accurate as needed by adding more terms and second, the virial coefficients can be calculated from theoretical models of the intermolecular potential energy functions. In order to understand these intermolecular potential functions we now turn our attention to the electric force.

7.4 The Electric Force Is the Essential Interaction that Leads to the Chemical Nature of the Universe

The essence of chemical interactions is electrical. Biophysical chemistry is built on the foundation of the physics of electricity. Electricity results from the separation of charge and is concerned with the behavior of charges; either at rest or in dynamic motion. The study of the behavior of charges at rest is called *electrostatics*. Electrodynamic phenomena occur when charges are in motion, and these are described by the laws of *electromagnetics*. The transport and behavior of charge in chemical systems are the basis of *electrochemical* studies. The type of charge carrier depends on the environment. In metal conductors charge is carried predominantly by electrons, whereas ions carry the majority of the charge in aqueous solutions. The study of interactions of ions with their environment is called *ionics*; the study of electron behavior in the solid state is *electronics*; and the study of charge transfer across an interface (an electrode) is the study of *electrodics*. We focus first on electrostatic behavior.

7.4.1 Electrostatics Define Electrical Forces Between Stationary Charges

If charges are stationary and are separated in space, then the system is an *electrostatic system* and is treated in the following fashion: Two charges, q_1 and

$q_{2,}$ are separated a distance r. An electrical force results. The charges feel the force between them as either attractive or repulsive. The force produced is described with the point of reference being q_1 and extends between q_1 and q_2. If the charge on both particles is the same, whether positive or negative, the force is repulsive. An attractive force results when the charges are of opposite sign. By convention, the force is said to be directed from the positive toward the negative body. The amount of charge carried by an electron (e) is considered to be the fundamental unit of charge, and its magnitude depends on the system of measurement. In SI units, the charge is measured in coulombs (C) and an electron carries approximately 1.602×10^{-19} C. In the older cgs system, the charge on an electron was an electrostatic unit (esu). Protons have an exactly equivalent but opposite charge compared to electrons Two important book-keeping rules related to charge are that

1) Charge is absolutely conserved.
2) Charge is quantized in integral multiples of e.

The magnitude of the force between charges with respect to their separation is described by *Coulomb's law*:

$$F = k\frac{q_1 q_2}{r^2} \tag{7.44}$$

The value of k depends on the units used. Coulomb's law as written in Eq. (7.44) has $k = 1$; q is given in esu and r in centimeters. The magnitude of the force described is in dynes. In the SI system, k is equal to $\frac{1}{4\pi\varepsilon_o}$, changing Eq. (7.44) into the following:

$$\mathbf{F}_{12} = \frac{q_1 q_2}{4\pi\varepsilon_o r^2}\hat{\mathbf{r}}_{12} \tag{7.45}$$

where ε_o is the permittivity of free space and has a value of 7.854×10^{-12} C^2/N/m^2; q is in coulombs, r in meters; and force is in N (newtons). The force is directional and is a vector. Therefore we have now added the unit vector, $\hat{\mathbf{r}}_{12}$ pointing between the two charges from q_1 to q_2.

Because electrical force is a vector, we would expect the principle of superposition to apply in cases of multiple point charges, which is the electrical many-body problem. This is experimentally verifiable. Solving multiple point charge problems is usually best done by drawing a clean descriptive diagram with coordinate axes and unit vectors. Searching for any symmetries in the system is valuable because it allows paired forces to be cancelled thus simplifying the calculations. The net sum of forces can then be found by calculating a vector sum:

$$\begin{aligned}\mathbf{F} &= \sum_{i=1}^{N} \mathbf{F}_{\mathrm{i}} \\ &= \frac{q}{4\pi\varepsilon_o}\sum_{i=1}^{N}\frac{q_{\mathrm{i}}}{r_{\mathrm{i}}^2}\hat{\mathbf{r}}_{\mathrm{i}}\end{aligned} \tag{7.46}$$

Table 7.4 Solutions to common distributions of charge, Q, sensed by a charge q

Distribution	Where q is	Force
Ring of charge (radius $= R$)	Located on axis, ($L=$ distance from center)	$\frac{qQ}{4\pi e_o}\frac{L}{\left(R^2+L^2\right)^{3/2}}$
Straight rod, aligned on x-axis, $L =$ length, center at origin	On x-axis at R from origin	$\frac{2ql_o}{4\pi e_o L}\left\{\ln\left[\frac{R-(L/2)}{R+(L/2)}\right]+ \mathrm{R}\left[\frac{1}{R-(L/2)}-\frac{1}{R+(L/2)}\right]\right\}\mathbf{i}$

We will consider cases in the next section in which large numbers of charges are considered together (Table 7.4). Under these conditions superposition becomes impractical, and the abstraction is reasonably made that it is of no physical consequence to consider the discreteness of charge. (*Caution*: This abstraction may need to be reconsidered in cases of biological relevance.) The approach is to consider a continuous distribution of charge composed of many infinitesimally small volume units ΔV, each containing a charge element Δq. The interaction between a reference point charge and the continuous distribution of charge is then explored. This interaction is shown in Fig. 7.6 and can be written as

$$\Delta\mathbf{F} = \frac{q}{4\pi\varepsilon_o}\frac{\Delta q}{r'^{\,2}}\hat{\mathbf{r}}' \qquad (7.47)$$

The charge in the volume element depends on the charge density, ρ, contained in the volume element. We write ρ as a function of the distance, r', between the charge

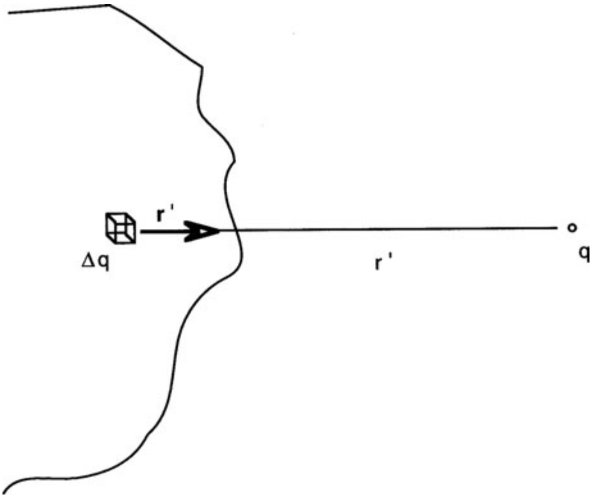


Fig. 7.6 Interaction between a point charge and a continuous charge distribution

element and q, i.e., $\rho\left(r'\right)$. The charge Δq contained in the volume element is then

$$\Delta q = \rho\left(\mathbf{r}'\right)\Delta V' \tag{7.48}$$

Combining (7.47) and (7.48) gives

$$\Delta \mathbf{F} = \frac{q}{4\pi\varepsilon_{\mathrm{o}}}\frac{\rho\left(\mathbf{r}'\right)}{r'^{\,2}}\hat{\mathbf{r}}'\Delta V' \tag{7.49}$$

As we cut each volume element into smaller and smaller pieces, the net force on q will be described by an integral:

$$\Delta \mathbf{F} = \frac{q}{4\pi\varepsilon_{\mathrm{o}}}\int\frac{\rho\left(\mathbf{r}'\right)}{r'^{\,2}}\hat{\mathbf{r}}'dV' \tag{7.50}$$

The integral may be easy to solve, and searches for symmetry can help provide a relatively straightforward answer. Alternatively, difficult integrations that cannot be solved analytically can be solved numerically with a computer.

In a vacuum the force that exists between two charged objects depends only on the distance of separation and the charges. If various non-conducting materials are inserted in the space between the charges, the force per unit charge decreases. These materials are called *dielectrics*, and the relative measure of their electrical-force attenuating properties is called the *dielectric constant* and is given the symbol ε. Since a vacuum has a dielectric constant equal to unity, the constants of all dielectric materials are greater than unity. Equation (7.45) is rewritten as

$$\mathbf{F}_{12} = \frac{q_1 q_2}{4\pi\varepsilon_{\mathrm{o}}\varepsilon r^2}\hat{\mathbf{r}}_{12} \tag{7.51}$$

7.4.2 The Electric Field Is Associated with a Charged Object

Earlier we mentioned the concept of a field as an abstraction that explains the effect of action at a distance. Anaximander ($\approx$ 550 BC) first proposed that there was an equilibrium of "actions at a distance", or forces, that reached out and held up the Earth. Michael Faraday suggested an alternative to the idea that a field emanated from one object, sensed a second, and then responded to the second body in a measured fashion. He proposed that a field could be thought to be an alteration in space induced by a single body. This field exists in space whether or not a second body is present. It becomes apparent when the second body is acted on by the field. Charges in space have an electric field associated with them.

An electrical potential field, E, is generated by the coulombic force described above. It is detected when it acts on a test object or charge, q_{o}, which is introduced at some distance, X, from the set of point charges:

$$\mathbf{E} = \frac{\mathbf{F}}{q_{\mathrm{o}}} \tag{7.52}$$

The electric field is a vector quantity of force per unit charge. In SI units, the field is in newtons/coulomb. Convention dictates that the field vector points in the direction toward which a positive test charge will be pushed. For a point charge, the electric field is found to be directed radially toward or away from the point charge depending on the polarity of the point charge, q_i, and the sensor charge, q_{o}. Thus the force on q_{o} due to q_{i} is

$$\mathbf{F}_{\mathrm{oi}} = \frac{q_{\mathrm{i}} q_{\mathrm{o}}}{4\pi \varepsilon_{\mathrm{o}} r^2} \hat{\mathbf{r}}_{\mathrm{io}} \tag{7.53}$$

and the electric field due to q_{i} is

$$\mathbf{E}_{\mathrm{i}} = \frac{q_{\mathrm{i}}}{4\pi \varepsilon_{\mathrm{o}} r^2} \hat{\mathbf{r}}_{\mathrm{io}} \tag{7.54}$$

The force on any charge in space is the charge times the field (any field) at that point in space:

$$\mathbf{F} = q\mathbf{E}_{\mathbf{ext}} \tag{7.55}$$

A useful facet of the field concept is that a field can carry energy. A field arising from a moving charge can propagate a field that travels at the speed of light. The electric field vector is known when its magnitude and direction is defined at every point in space. In other words, the field is a vector space in which each point in space is represented by a different vector. Fields are relatively easy to manipulate algebraically and are best suited for determining numerical information about electric forces. It is somewhat more difficult to grasp the vector space visually.

To help visualize the electric field, the idea of electric field lines (much like rays of light) can be utilized (Fig. 7.7). This idea was introduced in the mid-nineteenth century by Faraday who used the term “lines of force.” Electric field lines are an abstraction of the field itself and the lines are drawn as continuous in space. Two rules govern the drawing of a field line:

1) The tangent to a field line represents the *direction* of E at that point.
2) The density in space of the lines is proportional to the *magnitude* of the field at that point.

Field lines are drawn by convention as originating on positive charges and running outward from them. They terminate on negative charges. If the space is otherwise charge free, the lines continue on in straight line. The lines never cross one another because, if a crossing occurred, the direction of the electric field would be ambiguous at that point. The number of field lines and their density is set by making a

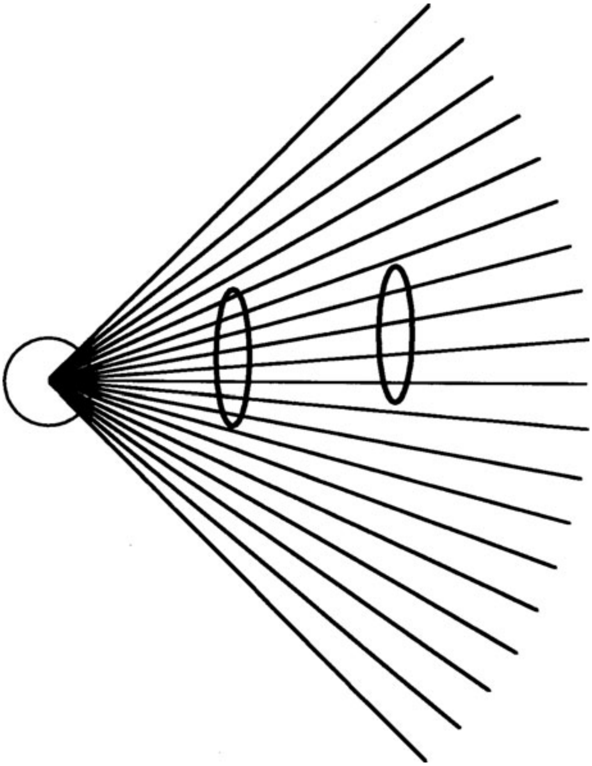


Fig. 7.7 The inverse square relationship can be seen geometrically by counting the number of force lines that pierce a surface as the distance from the origin changes

convenient choice of lines N that originate at a particular charge q and then the number of lines that originate or terminate on other charges is determined ratiometrically as $N_{\mathrm{i}} = \frac{q_{\mathrm{i}}}{q}N$. New field lines cannot be created in charge-free space.

Having noted these properties of field lines, the inverse-square relation of Eq. (7.54) can be derived as follows: Imagine a point charge with symmetrically drawn radial field lines directed away from the charge (Fig. 7.7). If a plane is drawn perpendicular to the direction of the field lines, in a small region the lines can be approximated as parallel to one another. The density of the field is the number of lines that pierces the plane and hence the density of the field is lines per unit area. If we drop a plane of unit area at r and then at $2r$ and count the number of lines piercing it, we will find the density of lines at in the plane at $2r$ to be $\frac{1}{4}$ the number compared with the plane at r which demonstrates the inverse square relationship. Figure 7.8 illustrates a variety of electric fields drawn with the lines-of-force abstraction. When considering this diagram and ones like it, remember this caveat: Visually compelling abstractions can often become more convincing then their formulation deserves. Always remember that *abstractions* are *practical tools* derived from simplifying assumptions.

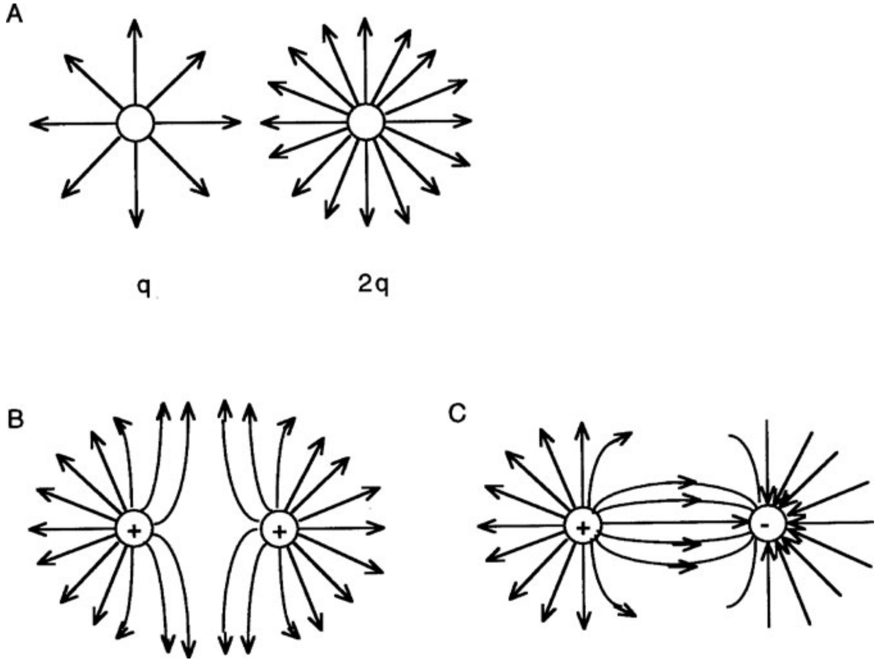


Fig. 7.8 Lines of force. (**a**) The field is represented by the number of lines of force and is proportional to the charge. (**b**) The lines of force between two similar charged bodies show the repulsive nature of the interaction. (**c**) The lines of force between two dissimilar charged bodies show the attractive nature of the interaction

Continuous collections of charge also give rise to an electric field. The electric field can be calculated by dividing the distributed charge into multiple infinitesimal elements whose distance from a reporter charge is r. The field due to each element is given by

$$d\mathbf{E} = \frac{dq}{4\pi \varepsilon_{\mathrm{o}} r^2}\hat{\mathbf{r}} \tag{7.56}$$

The total electric field is then found by integrating over the entire charge distribution:

$$\mathbf{E} = \int d\mathbf{E} = \frac{1}{4\pi \varepsilon_{\mathrm{o}}} \int \frac{dq}{r^2}\hat{\mathbf{r}} \tag{7.57}$$

For a linear-charge distribution of length L, the linear-charge density is given by $\lambda = \frac{q}{L}$and $dq = \lambda dx$. Similarly a planar or surface charge density over total area A, the surface charge density is given by $\sigma = \frac{q}{A}$, the charge over the differential area is $dq = \sigma dA$. Finally, the charge in a volume element depends on the volume charge density, $\rho = \frac{q}{V}$ and $dq = \rho dV$.

The electric field can be calculated using these tools for the field at distance R from an infinite sheet of charge with uniform surface charge σ. This solution will be important in a number of coming discussions and is given here as

$$\mathbf{E}_{\perp} = \frac{\sigma}{2\varepsilon_{\mathrm{o}}} \tag{7.58}$$

In (7.58) the electric field associated with an infinite sheet of charge is found perpendicular to the surface and constant in both magnitude and direction, the field does not even vary with the distance from the plane. In reality the infinite plane does not exist, and edge effects become important in the finite plane. The above description of the field will hold for a point found at distances much closer to the plane than the distance to the edge of the plane. Finally for two planes of opposite charge held parallel to one another the field is additive between the planes but cancels out and is zero everywhere else. Thus the field can be written as

$$\mathbf{E} = \frac{\sigma}{\varepsilon_{\mathrm{o}}} \tag{7.59}$$

This is the description of a parallel plate capacitor.

7.4.3 *Electric Dipoles Are Opposite Charges that Are Separated in Space*

An important case that often recurs in biophysical chemistry is the electrical dipole, which consists of two charges of opposite sign (q^{-}, q^{+}) separated by a distance, L. Many molecular species of concern in biochemistry, including water, are treated as electrical dipoles. Because the two charges do not sit one atop the other, their electric fields which would decrease as $\frac{1}{r^3}$ instead of $\frac{1}{r^2}$ This relationship holds when $r >> L$ as is discussed in Chapter 15. The electric field depends on the *electric dipole moment*, **p** where

$$\mathbf{p} = qL \tag{7.60}$$

p is a vector with the direction pointing from negative to positive charge. The electric field of the dipole along the bisecting axis of L is

$$\mathbf{E} = \frac{\mathbf{p}}{4\pi\varepsilon_{\mathrm{o}}r^3} \tag{7.61}$$

We will examine dipole interaction in much greater detail in Chapter 15.

7.4.4 The Electric Flux Is a Property of the Electric Field

Earlier we described the electric field as a series of vectors in space, with each point in space characterized by a vector. This is a general characteristic of any field: it is a physical quantity that has a different value at different points in space. Thus we have electric fields and magnetic fields as well as temperature fields (temperature is a scalar field, i.e., without direction.) The velocity field associated with a flowing liquid, i.e., *v(x,y,z,t)* represents the velocity of the liquid at each point in space at different times. The field lines are really tangents drawn to the vectors that exist at each point in space. A picture of the vector field is thus a function of position and time.

Vector fields have two mathematical properties that are helpful in understanding electromagnetism:

> The first is the idea that the vector field has flow. This is often described through the image of a flowing liquid. A flowing liquid is described by a velocity field and while we can easily relate to the physical movement of a flowing liquid, it is not the velocity field itself that is flowing. The abstraction of this "flow" is general to all fields including the electric field. We imagine a closed surface, a sphere, for example, and examine if the vectors all seem to move in one direction relative to the surface. Do all the velocity vectors move outward for example? Or we can ask more generally: Is the net flow outward across the surface (i.e., does more fluid flow outward than inward)? The net amount across the surface in a unit time is the flux of velocity through the surface. This flux is measured as the component of the velocity vectors *perpendicular* to the surface. So *flux* is the average normal component times the surface area. We denote flux as Φ.

> The second property of the vector field refers to its apparent flow along a loop or a line. We consider a flowing liquid (or rather its velocity field) and imagine that we place a tube of uniform internal diameter into the velocity field such that it loops around and closes on itself. We suddenly freeze all the liquid outside the tube only. The liquid will continue to move inside the tube in the direction of its net momentum. The *circulation* of the fluid in the tube will be the speed of the liquid times the circumference of the tube. Again more generally, the circulation is a quality of the vector field (nothing actually has to be moving), which describes the product of the average *tangential* component of the vector and the circumference of the closed loop.

These properties of the vector field describe the perpendicular and tangential components of the electromagnetic field and relate the electric (**E**) and magnetic (**B**) fields to one another for conditions of static or dynamic **E** and **B** fields. In the next sections we will relate how these descriptors of the vector field, flux and circulation, give rise to our descriptions of electrical and magnetic fields and hence to the fundamental basis of our understanding of the essential physics of chemistry and biology.

7.4.5 Gauss' Law Relates the Electric Field to an Electric Charge

The first law of electromagnetism is Gauss' law. This law relates the flux of an electric field through a closed surface to the charge contained by the surface. Gauss' law is a fundamental restatement of Coulomb's law. Their equivalence has been experimentally demonstrated to an extremely high degree of certainty. The flux through a closed surface that encloses no net charge must be zero since every line that enters the surface must also pass through and leave with the net result being zero. The law states that the flux is proportional to the total charge enclosed by the surface:

$$\Phi = \frac{q}{\varepsilon_o} \tag{7.62}$$

Gauss's law can easily be seen to be a more general statement of the Coulomb relation by imagining a point charge enclosed inside a spherically symmetrical surface of radius r and calculating the field strength (density of flux lines) at r and then recalculating when the closed surface is moved to $2r$. This is essentially the description of Fig. 7.7.

A number of important conclusions can be derived from Gauss' Law and they are summarized here.

1. *Inverse square relationship*: As described above, the $\frac{1}{r^2}$ relationship of field strength to distance can be directly found.
2. *Charge in a conductor*: The electric field on the inside of a conductor in electrostatic equilibrium is zero.
   > The excess electrostatic charge on a conductor resides on its surface, there is no net charge in the interior of the conductor.
   > The magnitude of the electric field at any point just outside a conductor is proportional to the density of the surface charge at that point, i.e.

$$\mathbf{E} = \frac{\sigma}{\varepsilon_o} \tag{7.63}$$

3. *Cylindrical symmetry*: A long straight wire with a positive uniform linear-charge density (λ) has an electric field that extends radially away from the wire and is equal at each point, r, equidistant from the wire. Over the rounded portion$\int \mathbf{E}dA = 2\pi \mathbf{E}rl$, and the field at the ends is 0. The enclosed charge is $q = \lambda l$ and the electric field is

$$\mathbf{E} = \frac{1}{2\pi\varepsilon_o r} \tag{7.64}$$

4. *Spherical symmetry*: A Gaussian surface concentrically encloses a *spherical shell* of symmetrical charge, Q. The flux through the surface is $\Phi = 4\pi r^2\mathbf{E}$ where r is the radius of the Gaussian surface.

$$\mathbf{E} = \frac{Q}{4\pi\varepsilon_o r^2} \tag{7.65}$$

- For a *uniform sphere of charge* of radius R, (this sphere is not a conductor), the field inside the sphere at a distance r from the center is $\mathbf{E} = Q'/4\pi\varepsilon_o r^2$ with Q' the charge inside of r which is found by considering the charge density of the volume, ρ. $Q' = 4\pi/3\rho r^3$ and $\mathbf{E} = \rho r/3\varepsilon_o$. In the case where the total charge is Q, ρ will be $3Q/4\pi r^3$ and it follows that

$$\mathbf{E} = \frac{Qr}{4\pi\varepsilon_o R^3} \tag{7.66}$$

5. *Planar symmetry*: A sheet of charge with a surface charge density, σ, has a field perpendicular to the surface with uniform intensity,

$$\mathbf{E} = \frac{\sigma}{2\varepsilon_o} \tag{7.67}$$

Gauss' Law can be used to find electric field when a charge distribution is given, especially when a high degree of symmetry is present. Many of these examples are worked out in standard engineering physics textbooks (see reference list) and should be examined if the student desires a more intimate knowledge of the subject.

7.4.6 A Point Charge Will Accelerate in an Electric Field

What happens if a small charged particle is placed in an electric field? The force can be described in terms of Newton's second law:

$$\mathbf{F} = \mathrm{q}_{\mathrm{ext}} = m\mathbf{a} \tag{7.68}$$

Example

For an electron with mass of 9.11×10^{-31} kg placed in a field generated by a pair of two oppositely charged parallel plates with $\sigma = 1.0 \times 10^{-6}$ C/m^2 and separated by 1 cm, the acceleration caused by the electric field will be

$$\begin{aligned}\mathbf{a} &= \frac{q_{ext}}{m} = \frac{q_{ext}\mathbf{E}}{m} = \frac{q_{ext}\sigma}{m\varepsilon_o} \\ &= \frac{(1.6 \times 10^{-19}\mathrm{C})(1.0 \times 10^{-6}\ \mathrm{C/m^2})}{(9.11 \times 10^{-31}\ \mathrm{kg})(7.85 \times 10^{-12}\mathrm{C^2/N\text{-}m^2})} \\ &= 1.9845 \times 10^{16}\mathrm{m/s^2}\end{aligned} \tag{7.69}$$

Since the electron can travel 1 cm from one plate before striking the other plate, it will attain a velocity of

$$\begin{aligned} v^2 &= 2ax \\ v &= \left[2(1.9845 \times 10^{16}\text{m/s}^2)\left(1 \times 10^{-2}m\right)\right]^{1/2} \\ &= 1.992{\times}10^{7}\text{m/s} \end{aligned} \tag{7.70}$$

and will take

$$\begin{aligned} t = \frac{v}{a} &= \frac{1.992 \ \times \ 10^{7} \text{ m/s}}{1.9845 \ \times \ 10^{16} \text{ m/s}^2} \\ &=1.003 \times 10^{-9}\text{s} \end{aligned} \tag{7.71}$$

to travel that distance.

Similar treatments can be applied to the effect of the electric field on a moving particle by examining the effect of the field vector on the appropriate component of the traveling particle's motion. As the above example suggests the velocity of a particle can be substantial, and as the velocity approaches the speed of light, relativistic effects must be taken into account.

7.4.7 The Electric Potential Is the Capacity to Do Electrical Work

The electric or Coulomb force is a conservative force like gravity. This force is a fundamental relationship in nature. Because the force is conserved, collected charge will have a potential energy that can manifest as a kinetic energy under the proper conditions. While electric forces involve the interaction between a distribution of charge and a second test charge, we have described the effect of the electric force around an isolated charge distribution through the abstraction of the electric field. The electric force is the product of a test charge and the field. The electric potential energy is the energy of the charge distribution and the second charge. The magnitude of the electric field is a derivative of the electrical potential, which is a scalar quantity.

An energy of position arises when an object that can be acted upon is placed in a force field. Any force that is a function of position alone is a conservative force; consequently an object that is influenced by the force will have a potential energy associated with it. The potential energy (U) can be converted to kinetic energy (KE) as allowed by the law of conservation of energy. The total change in energy $\Delta E = 0$ and $\Delta U = -\Delta KE$. Because potential energy is a function of position, a change in potential energy means that there has been a change in position in the vector field. Only changes in potential energy have any physical meaning. (This is why a squirrel can run unaffected along a high-voltage power line, it remains on an isopotential surface). The potential energy change is associated with the movement of the object from position r_1 to r_2 along a displacement, s

$$\begin{aligned}\Delta U &= U_2 - U_1 = U(r_2) - U(r_1) \\ &= -\int_{\mathbf{r}_1}^{\mathbf{r}_2} \mathbf{F} d\mathbf{s}\end{aligned} \tag{7.72}$$

Conservative forces are characterized by a line integral whose value is independent of the path of integration taken for the displacement.

The change in electric potential energy for a system with a point charge, q, at the origin and a test charge that is also a point charge, q_o, that is moved from a to b can be determined from Eq. (7.72). Because this is a system with radial displacement we can write $\mathbf{F}\cdot d\mathbf{s}$ in terms of $\mathbf{F}\cdot dr$:

$$\begin{aligned}\Delta U &= -\int_{r_a}^{r_b} F \cdot dr = -\int_{r_a}^{r_b} \frac{qq_o}{4\pi\varepsilon_o r^2} dr \\ &= -\frac{qq_o}{4\pi\varepsilon_o}\int_{r_a}^{r_b} \frac{dr}{r^2} \\ &= -\frac{qq_o}{4\pi\varepsilon_o}\left(\frac{1}{r_b} - \frac{1}{r_a}\right)\end{aligned} \tag{7.73}$$

This solution is general and will hold whether movement of the test charge is along the same or different radial paths. Since physical consequences are only associated with changes in potential energy we are free to choose a reference point that is convenient. The electric potential energy function is chosen to be zero at infinity because at infinity the interaction energy actually is zero. The electric potential energy $U(r)$ for a system of point charges separated by a distance r is

$$U(r) = \frac{qq_o}{4\pi\varepsilon_o}\frac{1}{r} \tag{7.74}$$

The physical meaning of the potential energy is that it represents the amount of work needed to bring one charge from infinity to point of finite separation between the charges.

Accordingly, we see that a charge generates an electric field that exists in the surrounding space. Through this field it exerts a force on any charge that moves into that space as $\mathbf{F} = q_o\mathbf{E}$**.** The charge also creates the capacity for a potential energy to become associated with the system upon the introduction of a charge q_o at some distance from the charge. This potential energy *U(r)* can be written in a fashion analogous to the force equation:

$$U(r) = q_o V(r) \tag{7.75}$$

This means that the charge, q, generates an electric potential, $V(r)$, which will affect any charge q_o that is brought to r. If this relationship is to hold as written, the test charge should be small enough so it does not disturb the charge distribution giving rise to the potential. The electric potential is written in terms of work per unit charge or joules/coulomb. Like the electric field, $V(r)$ is independent of the test charge.

The electric potential can be found for a point charge by combining Eqs. (7.74) and (7.75):

$$V(r) = \frac{U(r)}{q_{\mathrm{o}}} = \frac{q}{4\pi\varepsilon_{\mathrm{o}}r} \tag{7.76}$$

The electric potential difference between two points will be

$$\begin{aligned} \Delta V = V_b - V_a = \frac{U(r_b) - U(r_a)}{q_{\mathrm{o}}} \\ = \frac{q}{4\pi\varepsilon_{\mathrm{o}}r}\left(\frac{1}{r_b} - \frac{1}{r_a}\right) \end{aligned} \tag{7.77}$$

We mentioned that the electric potential is a scalar quantity in contrast to the electric field, which is a vector quantity. Both are created by a distribution of charge, but it is generally simpler to work with scalar quantities. The electrical potential difference can be seen as an integral over the electric field by rewriting the second term of Eq. (7.77) in the form of (7.72) and remembering that $\mathbf{F} = q_o\mathbf{E}$:

$$\begin{aligned} \Delta V = \frac{U(r_b) - U(r_a)}{q_{\mathrm{o}}} \\ = -\int_{r_a}^{r_b} \mathbf{E} \cdot d\mathbf{s} \end{aligned} \tag{7.78}$$

This is a general expression for the potential difference between two points and is expressed as a path-independent integral over the electric field. The change in potential energy in a system is equal to the negative of the work done by the system to move a test charge from *a* to *b*. Thus the physical meaning of Eq. (7.78) is that ΔV is the work per unit charge needed to move a test charge from *a* to *b* without changing the kinetic energy of the system.

In SI units the electric potential is given in joules per coulomb (J/C) since it represents the energy per charge. For convenience the unit of 1 J/C is called the *volt*. The electric field can be described as force per charge (N/C) or perhaps more commonly as volts per meter (V/m). This relationship is found in (7.78). Though not an SI unit a commonly used term is the *electron-volt* (eV) which represents the charge of an electron times the voltage. The electron-volt is a unit of energy and is related to the joule as follows:

$$1\,\mathrm{eV} = \left(1.6 \times 10^{-19}\mathrm{C}\right)(1\,\mathrm{V}) = 1.6 \times 10^{-19}\mathrm{C}\,(\mathrm{J/C}) = 1.6 \times 10^{-19}\mathrm{J}$$

Like the electric field, the electric potential obeys the superposition principle. The potential due to a collection of charges is simply the *scalar* sum of the charges. For a continuous distribution of charge the voltage is given by

$$V = \int dV = \frac{1}{4\pi\varepsilon_{\mathrm{o}}}\int \frac{dq}{r} \tag{7.79}$$

7.4.8 *Equipotential Surfaces Are Comprised of Lines of Constant Potential*

Let us think about that squirrel running along a high tension wire. This squirrel clearly has significant potential energy both in terms of its gravitational potential energy and in terms of its electric potential energy. The former is determined by its height and the latter by the charge carried in the wire. If the squirrel were to step off the wire or contact a grounded object (electrical potential = 0), it would quickly, albeit briefly recognize its potential energy. The squirrel runs blithely along with no significant consequence or awareness of its potential danger. Since only changes in potential energy are of physical consequence, the squirrel is safe as long as it runs along a line of constant potential energy. Such lines are called *equipotential lines* or surfaces. If these points of constant potential are connected a potential energy map can be drawn, much like a topographer's map. The squirrel can run along the equipotential line that the high tension wire constitutes with no adverse effect because its potential energy remains the same along the line scribed by the wire. Obviously neither the force of gravity nor the electric force has a component along the equipotential line. However, if the squirrel steps off the line, it will discover that both the gravitational and the electric forces exist in a direction perpendicular to the equipotential line, and the squirrel will therefore accelerate in a direction perpendicular to the line. All conservative forces have these characteristics. Because there is no force acting along the equipotential line, no work is done on an object by that force. The object moves freely along the equipotential line or surface.

Because the electric force is everywhere perpendicular to the equipotential surface, the field can be easily found if the equipotential surfaces are known. Conversely the electric potential contour map can be determined from the electric field. We already know that the radial electric field of a point charge will generate a series of equipotential concentric spherical surfaces around the charge decreasing as $\frac{1}{r}$. In addition, we discovered earlier that the electric field between two parallel plates of very large dimension is perpendicular to the plates; accordingly, the equipotential surfaces will be a series of surfaces parallel to the charged plates with a linear change in potential with respect to the distance away from the charged plate. These examples are illustrated in Fig. 7.9. The electric field is found to point along the sharpest gradient of the equipotential lines, or in other words, along the shortest distance between equipotential lines.

7.4.9 *Calculating Potential Fields*

Implicit in much of our work in biophysical chemistry is the need to consider the contribution of the potential energy of the electric field. We will often see how the electric field influences systems of biological importance. A working knowledge of how to find equipotential surfaces is a valuable tool for the serious biological worker. A qualitative picture of the potential fields can be relatively easily drawn

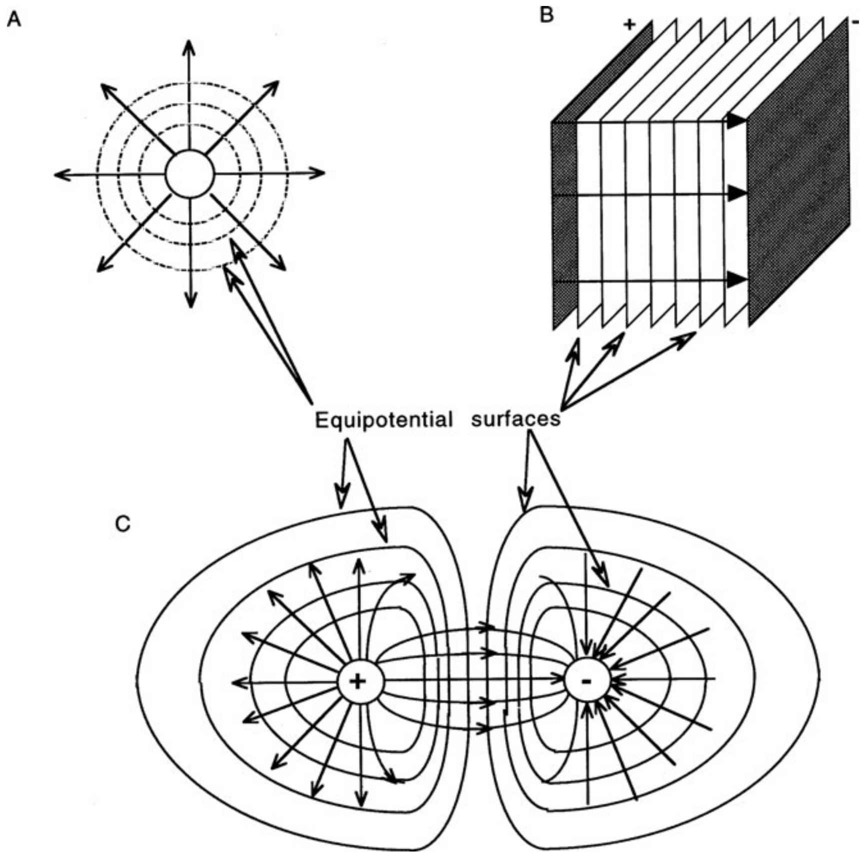


Fig. 7.9 The electric field and equipotential surfaces for a point charge, parallel plates of charge, and a dipole

by using graphical methods; but if quantitative methods are required we will either derive the potential surfaces from a knowledge of the electric field, $-\int_{r_a}^{r_b} \mathbf{E} \cdot d\mathbf{s}$, or we can use the potential calculated from the following list:

$$\text{for a point charge } V = \frac{q}{4\pi\varepsilon_{\text{o}} r} \tag{7.80}$$

$$\text{for many point charges } V = \frac{q}{4\pi\varepsilon_{\text{o}}} \sum_{\text{i}} \frac{q_{\text{i}}}{r} \tag{7.81}$$

$$\text{for a continuous charge distribution } V = \frac{q}{4\pi\varepsilon_{\text{o}}} \int_{\text{i}} \frac{dq}{r} \tag{7.82}$$

Remember that a zero potential point must always be chosen. Implicit in the above equations is the convention that the potential is zero at infinity, and therefore, the

Table 7.5 Electric fields and potentials for charge geometries

Geometry	Field magnitude	Potential	Point of zero *potential*
Point charge	$\frac{q}{4\pi e_o r^2}$	$\frac{q}{4\pi e_o r}$	∞
Parallel plates (infinite dimension, oppositely charged, uniform charge density $= \sigma$, separation $= d$)	$\frac{s}{e_o}$	$-\frac{sd}{e_o}$	Anywhere
Electric dipole (dipole moment $= p$,)	$\frac{p}{4\pi e_o r^3}$ along bisecting axis, far away	$\frac{p \cos q}{4\pi e_o r^2}$ far away but everywhere	∞
Charged ring (radius $=$ R, along axis at distance x)			∞
Uniformly charged sphere	$r \geq R : \frac{Q}{4\pi e_o r^2}$	$r \geq R : \frac{Q}{4\pi e_o r}$	
(non-conducting and solid, radius $= R$)	$r \leq R : \frac{Q}{4\pi e_o R^3}$	$r \leq R : \frac{Q}{8\pi e_o}\left(3 - \frac{r^2}{R^2}\right)$	∞

choice has already been made. The equations for the electric field and electric potential in a variety of common geometries are summarized in Table 7.5. For the general case see Appendix F.

7.4.10 Capacitors Store Electrostatic Field Energy

We have seen that two conducting plates may be charged and that a potential field can be found to exist between them. This potential field is capable of doing work. This was demonstrated earlier when we considered the acceleration of a charge placed between the plates. Such a storage device is a *capacitor*, and it stores energy by storing charge. The relationship between charge and potential is called the *capacitance*:

$$C = \frac{Q}{V} \tag{7.83}$$

where Q is the charge in coulombs, C is the capacitance in farads $\left(1\ \mathrm{F} = \frac{1\ \mathrm{Coulomb}}{1\ \mathrm{Volt}}\right)$, and V is the potential in volts across the capacitor. In a parallel plate capacitor, the capacitance is given by the equation:

$$C = \frac{\varepsilon\varepsilon_0 A}{d} \tag{7.84}$$

where A is the area of the plates, ε is the dielectric constant of the material between the plates, and d is the distance between the plates. (Other physical arrangements have different expressions for the capacitance, for example, for a conducting sphere, $C = 4\pi\varepsilon_0 r$ where r is the radius of the sphere.) Capacitance is always positive and has units of coulombs per volt (farads). Commonly smaller units of capacitance such as the microfarad or picofarad are practically encountered. Capacitors do not allow current to flow through them because the material separating the plates, called the *dielectric*, is electrically insulating. When an external voltage source is applied across a capacitor, charge is moved from the battery to the plates. The plate connected to the negative side of the voltage source, build up excess of electrons and the plate becomes negatively charged. Conversely, the plate connected to the positive side of the battery develops a positive charge as electrons are withdrawn from it. As the charges build up on the plates, an electric force field is generated that ultimately is exactly the same as the voltage of the external source. However, the field direction across the capacitor is opposite to that of the external source. The potential difference falls to zero because of the counterpotential; a point is therefore reached when no more charge can be added and the system is in equilibrium. What happens when a dielectric material is inserted between the charged plates? Inevitably, the measured potential field across the capacitor diminishes, in a degree depending on the dielectric material chosen. This occurs without the loss of charge since the same charge is still on the plates. The capacitance of the parallel plate capacitor increases solely because the dielectric changed. The dielectric constant is found by comparing the capacitance of a parallel plate capacitor with a dielectric of vacuum versus any other material.

$$\varepsilon = \frac{C_{\text{any dielectric}}}{C_{\text{vacuum}}} \tag{7.85}$$

We will explore the molecular basis of the dielectric effect in Chapter 15.

Energy is stored in capacitors when they are charged. The energy in a capacitor can be found by calculating the work done to bring one charge from one plate to the other thus creating a charge separation of $+dq$ on one plate and $-dq$ on the other. The work needed to move the second charge will be greater than that needed to move the first. The next charge will be abstracted from a plate that has unlike charge and deposited on a plate with like charge. With each iterative separation of charge a greater amount of work will be required. The potential difference created by the separation of each charge will oppose the transfer of the next charge. Therefore we can write

$$dW = Vdq = \frac{q}{C}dq \tag{7.86}$$

The total work required starting with zero charge and ending up with $\pm Q$ is

$$W = \int dW = \int_0^q \frac{q}{C}dq = \frac{1}{C}\int_0^q q\,dq = \frac{Q^2}{2C} \tag{7.87}$$

The work done is stored in the capacitor as potential energy and so Eq. (7.87) can be written as

$$W = U = \frac{Q^2}{2C} \tag{7.88}$$

Using the relationship $Q = CV$, we can determine the potential energy stored if we know the total charge, the potential, or both

$$W = U = \frac{Q^2}{2C} = \frac{CV^2}{2} = \frac{QV}{2} \tag{7.89}$$

The energy in a capacitor is stored in the electric field itself, and the energy density can be described as the energy per unit volume, U/V. In electrical systems, two of the most common sources of energy storage are capacitors and batteries. Batteries store energy in chemical bonds, and this energy can be released by electrochemical reactions at the electrodes. The potential energy stored in a capacitor or battery can be released in a measured manner through an electric circuit so that the flow of charge down its potential gradient can be used to do useful work (as determined by the circuit elements). The flow of charge in a given unit of time is called *electrical current* its flow in electrochemical systems will be explored in later chapters.

7.5 Wave Motion Is Important in Electromagnetic and Mechanical Interactions in Biological Systems

Why do we need knowledge of wave motion? We know that the description of our natural state space has qualities of both particle mechanics and wave mechanics. When we require quantum mechanical descriptors we will need to be familiar with both of these mechanical descriptions. This knowledge is necessary for work at a fairly deep level. At a very practical level; however, the wave description per se is very useful abstraction. We know about the biological systems that interest us because we interact with them. Many of our interactions are through electromagnetic phenomena. Though all of these phenomena can be completely described in terms of quantum electrodynamics, this is practically (though not conceptually) complicated. Electromagnetic radiation is composed of rapidly alternating electric and magnetic fields, i.e., waves. It is the rapid to and fro motion of electrons that gives rise to the electromagnetic radiation that influences the electrons in our eyes and instruments and allows us to see and enables our instruments to sense. Without the connection of the wave, we would literally be blind to the world around us.

All of us have experience with wave motion. Therefore it is relatively easy to understand many aspects of biophysical studies by talking in terms of wave motion. We present this chapter as a useful background to x-ray diffraction, microscopy, light scattering studies including Raman spectroscopy, polarization studies, circular dichroism, and quantum mechanics.

Waves carry energy and momentum through a state space without the concurrent transport of matter. Mechanical waves such as water waves, sound waves, and waves on a spring, carry energy and momentum by propagating a disturbance in the medium. This occurs because of the intrinsic elastic properties of the medium. The energy and momentum of electromagnetic waves are carried by electric and magnetic fields that propagate through a vacuum. Much of our understanding of wave behavior including electromagnetic waves is based on our experience and knowledge of mechanical waves. Since mechanical waves propagate via a distortion of their medium, it will be easy to see why an “etheric medium” through which electromagnetic waves propagate was such an appealing, though incorrect, concept of classical science.

7.5.1 Pulses Are the Starting Point for Understanding Wave Motion

First we consider a *wave pulse*. A string is stretched placing it under tension. A single flip of the string generates an impulse perpendicular to its dimension of stretch. A bump in the string is produced by the impulse. The propagation of this disturbance down the string is a *wave pulse*. The speed of the pulse depends on the physical nature and tension of the string. The wave pulse is a distortion in the equilibrium shape of the string and has a definite shape as it begins to propagate. If we attach a weight to the middle of the string, we will see the weight rise and fall as the pulse reaches and passes the weight. This demonstrates that the wave front carries both energy and momentum provided by the work done in lifting the string initially. As the pulse travels, the shape changes gradually, spreading out. This process of spreading is called *dispersion* and is a property common to all waves, except electromagnetic waves that propagate in a *vacuum*.

We will defer discussion of dispersion for now and deal with the idealized nature of a wave pulse. First of all, the principles of the pulse are quite general and apply to a flash of light (lightning), a pulse of sound (a crash of thunder), or a pulse of mechanical energy (a tidal wave). The pulse carries the transient disturbance in the shape of the medium but the medium itself is not transmitted. In the case of the pulse on the string, the mass elements of the string are moving perpendicular to the direction of the wave propagation, therefore this is a *transverse wave*. Electromagnetic radiation, comprised of electrical and magnetic fields whose wave vectors vibrate perpendicular to the direction of wave propagation, are transverse waves. In contrast, sound waves cause a disturbance in their medium that is in the direction of wave propagation. The vibration of a string or a tuning fork causes a to and fro motion of the air molecules which alternately compresses and decompresses the air; thus the shape disturbance propagates in a series of parallel fronts. Such waves are named *longitudinal waves*. A water wave is a combination of transverse and longitudinal waves with the local motion of the molecules in a circular path (Fig. 7.10).

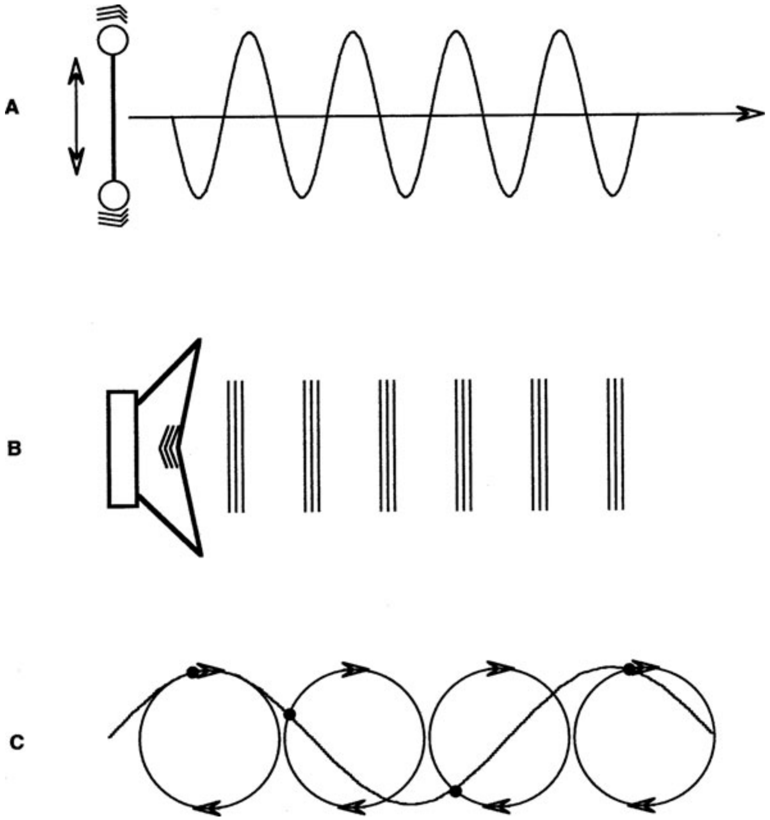


Fig. 7.10 Wave types: (**a**) transverse waves, (**b**) longtitudinal wave, (**c**) water waves (a combination of transverse and longitudinal waves)

7.5.2 *The Wavefunction Is a Mathematical Expression for Wave Motion in Terms of Space and Time*

We will consider a one-dimensional line first: A disturbance in shape of the medium can be written as a function in terms of time and space coordinates. For a transverse wave such as the pulse on a string, the shape of the pulse at $t = 0$ is $y = f(x)$. Neglecting dispersion so that $f(x)$ does not vary with time, a frame of reference can be defined that moves with a velocity, v, such that the shape is always $y' = f(x')$. The two reference frames are related

$$y = y' \tag{7.90}$$

and

$$x = x' + vt \tag{7.91}$$

A pulse moving to the right will be represented by

$$y = f\,(x - vt) \tag{7.92}$$

and a wave moving left can be written as

$$y = f\,(x + vt) \tag{7.93}$$

The function, $y = f\,(x - vt)$, is called the *wavefunction* and describes the vertical displacement in terms of time, t, and position, x. v is the velocity of the wave pulse. Analogous wavefunctions can be written for waves of different types; for example, the electromagnetic waves can be written in terms of its electric and magnetic vectors, $E\,(x - vt)$ and $B\,(x - vt)$, respectively.

7.5.3 *Superposition and Interference Are Fundamental Properties of Wave Interaction*

Waves seldom exist in isolation, and when more than two waves are present it is important to be able to describe the state of the system as well as the interaction of the waves. The wavefunction is extremely valuable in this application. Consider two wave pulses on the string, which are moving toward one another. Upon meeting, the shape of the resultant pulse is the sum of each individual wave added together. Figure 7.11 shows three cases of pulses meeting on the string.

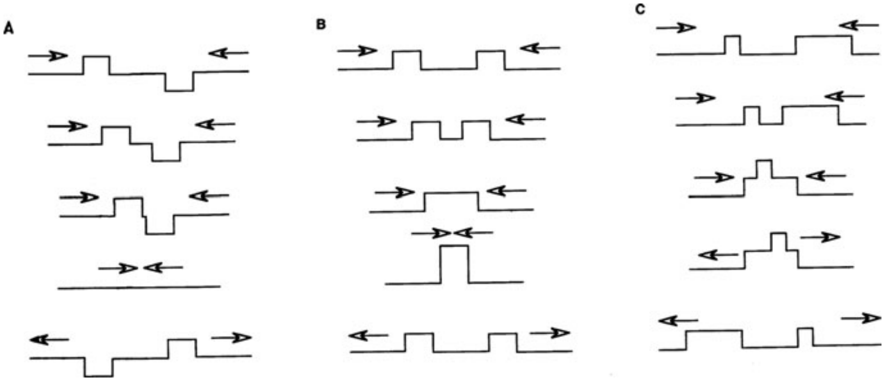


Fig. 7.11 Interaction of waves meeting on a string are fundamental qualities of waves

In the first case two equal but inverted pulses meet. At the moment just before their maximums meet, each pulse is identifiable and appears to be a mirror image of the other. At the moment of their meeting there is no movement of the string, and the pulses add to zero. Though the pulses in this system cannot be seen or located, *the string is not at rest*. An instant later, the waves reappear each continuing in their own course.

In the second case, two equal waves approach; as in the case above the waves are identifiable before and after they meet. However, at the moment when they are superimposed upon each other, the resultant wave has high amplitude but an equal period to the original pulses.

The analysis of the third more general case is left to the reader.

The combination of separate waves to produce a resultant wave is a unique property of wave motion and is called *interference*. Interference is not seen in particle motion because two particles cannot overlap and be added together in this fashion. The mathematical adding together of the wavefunctions to get the resultant is called *superpositioning.* The *principle of superposition* states that a resultant wavefunction is an algebraic summation of the individual wavefunctions. Thus in the case above, the total wavefunction, $y\ (x, t)$ can be written as

$$y(x,t) = y_1(x - vt) + y_2(x + vt) \tag{7.94}$$

The special case given in the example, at the time of overlap, $t = t_1$:

$$y_1(x - vt) = -y_2(x + vt) \text{ and } \quad y(x,t) = 0 \tag{7.95}$$

When the amplitude of the resultant wave is greater than the original waves the interference is called *constructive* (Fig. 7.11b); when the resultant is smaller the interference is *destructive* (Fig. 7.11a).

The principle of superposition applies to all electromagnetic waves traveling in a vacuum. In general it is applicable to the motion of pulses that are not too large. In large amplitude cases, the interaction of the waves is not algebraic or linear. *Non-linear* waves and non-linear phenomena are extremely important but are quite complicated to deal with mathematically. It is often easier to deal formally with systems in the limit of their linear behavior yet many of the phenomena of biological and chemical interest become of interest when they become non-linear! In this book we will emphasize the linear formalisms as much as is useful for grasping the basic ideas

7.5.4 The Velocity of a Wave Pulse Is a Function of the Transmission Medium

The velocity of a wave depends only on the physical properties of the medium in which it travels and is not dependent on the motion of the source with respect to the medium. It bears repeating that electromagnetic waves require no medium for propagation, and their velocity (approximately 3×10^8 m/s) is at maximum in a vacuum. The velocity, v, of the pulse on the string above is dependent on the tension, T, in the string and the mass per unit length, μ, of the string. The relationship is

$$v = \sqrt{\frac{T}{\mu}} \tag{7.96}$$

7.5.5 Reflection and Transmission of a Wave Depends on the Interface Between Two Phases of Different Speeds of Propagation

What happens to the pulse when the end of the wave is reached? The fate of the pulse depends on the interface at the end of the string. Part of the pulse will be reflected and part will be transmitted depending on the speed of propagation in the subsequent material. The general rule is that when a wave pulse enters a region with a slower speed of propagation, the reflected pulse will be inverted. If the region has a higher speed of propagation, the reflected pulse is not inverted. If the object is essentially infinitely more dense (like an immobile wall), the pulse will be reflected but inverted. If the subsequent medium is a lighter string, part of the pulse will be reflected and not inverted, while part of the pulse will continue down the lighter string. Strings of equal density will transmit the pulse without change; a pulse traveling from light to heavy string will have the reflected pulse inverted and the transmitted pulse continue.

7.6 Harmonic Waves Are the Result of a Sinusoidal Oscillation

Suppose we hook a string to a tuning fork or an oscillating cam on a motor. The wave pattern generated is illustrated in Fig. 7.12. This is a special case of wavefunction called a *harmonic wave*. This wavefunction is quite important both practically and theoretically and will be discussed in some detail. An unusual feature of harmonic waves is that unlike the wave pulse discussed earlier, their shape does not change with dispersion in the propagating medium. If the momentum and energy of a harmonic wave is adsorbed in even a highly dispersive medium, the wave may decrease its amplitude but it will not change its shape. This is quite important in the mathematical description of a wave since the velocity of a wave usually needs to be defined. The velocity is found relating the time and distance traveled by a particular point on the wave, but if a wave is changing shape due to dispersion, which point

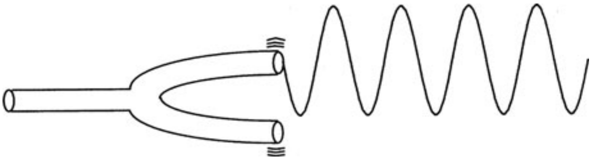


Fig. 7.12 Production of a harmonic wave on a string by a vibrating tuning fork

on the wave can be seen as typical of the speed of the wave? With a harmonic wave, any point can be chosen to represent the velocity of the wave. In fact it can be shown that the velocity can be precisely defined only for a harmonic wave.

Other waves and pulses that change shape when propagating in a dispersive medium can each be resolved into particular groups of harmonic waves which are summed by superposition. Thus a detailed treatment of dispersive waves is carried out on the constitutive harmonic waves. The mathematical technique of converting any wave to its harmonic elements (and all waves can be transformed in this way) is called *Fourier analysis*.

Harmonic waves have wavefunctions that are sine or cosine functions, i.e.,

$$y(x,t) = y_{\circ} \sin k(x - vt) \tag{7.97}$$

y_0 is the amplitude and k is the wave number. The reader may be more familiar with this function when written in terms of the angular frequency ωwhich is $\omega = kv$.

$$y(x,t) = y_{\circ} \sin(kx - \omega t) \tag{7.98}$$

7.6.1 Wavelength, Frequency, and Velocity

The distance in space separating two successive crests is called the *wavelength*, **λ**. Note that ωt is a *phase constant*, δ thus (7.98) can be written as

$$y(x,t) = y_{\circ} \sin(kx - \delta) \tag{7.99}$$

Using (7.99), the relationship between the wave number and wavelength can be found. The position of a crest is x_1 and that of the next is x_2. The distance $x_2 - x_1 = \lambda$. Thus within a single wavelength, the wave which is essentially described as a function of the rotation around a circle is repeated or makes a full revolution. The argument of $\sin(kx - \delta)$ changes by 2π. Thus

$$k(x_2 - x_1) = 2\pi \tag{7.100}$$

and

$$k\lambda = 2\pi \text{ or } k = \frac{2\pi}{\lambda} \text{ and } \lambda = \frac{2\pi}{k} \tag{7.101}$$

The frequency, f, and period, T are related to the angular momentum:

$$T = \frac{2\pi}{\omega} \text{ and } f = \frac{1}{T} = \frac{\omega}{2\pi} \tag{7.102}$$

The velocity of the wave is easily related to the frequency and wavelength:

$$v = f\lambda \tag{7.103}$$

7.6.2 Polarization

Let us now revisit some of the properties of wave motion that we discussed earlier in considering pulses. Transverse waves have their wave motion perpendicular to the direction of propagation and can be resolved into rectangular coordinates in the plane perpendicular to the direction of propagation (Keep in mind that we are talking here about one-dimensional wave motion. This subject will become more complex as well as more interesting when two and three dimensions are considered). Suppose a harmonic wave is traveling down the x axis. The vibration is then in the xy plane. A wave traveling in one direction in which all of the wave motions remain in the same plane is said to be *plane polarized* or sometimes *linearly polarized*. If the wave motion were in the xz plane rather than the xy plane, Eq. (7.97) would be written as

$$z(x,t) = z_{\circ} \sin k(x - vt) \tag{7.104}$$

We now consider plane polarized waves in three dimensions. Looking again at a plane polarized wave we see that it propagates in one direction (x), varies in amplitude in a plane perpendicular to the direction of propagation (and this can be any plane that includes the x axis), and has a definite phase relationship as it propagates. We will pick two plane polarized waves, both propagating in the x direction, one with vibrations in the xy plane and the second with vibrations in the xz plane, but both waves in phase. First we consider the case in which both of the waves are of equal amplitude. As these two waves propagate we can see that a new resultant wave can be found by superposition. To observe the resultants we will stand at the end of the x axis and watch as the wave comes toward us. The plane in which this new wave is oscillating will be seen to be 45° off the x axis or $\frac{\pi}{4}$ radians. If only the amplitudes of these two waves are varied, the resultant waves will rotate around the x axis as in Fig. 7.13a. Note in each case that the resultant wave remains plane polarized.

Now consider the case where we hold the amplitude of the two waves to be equal but allow the phase relationship of the two waves to vary (Fig. 7.14). Again we stand in the same place to observe and now we see that the resultant waves are linear, elliptical, or circular depending on the difference in the phase between the waves. When the phase difference is 0° or any integral multiple of $\frac{\pi}{2}$, the resultant wave is linear and the wave is *plane polarized*. If the phase difference is 90° or any odd integral multiple of , the resultant wave (when viewed on end) is circular and is said to be *circularly polarized*. These two cases are actually special cases of the more general form that the resultant wave takes, which is seen to travel around the x axis in an ellipse and is called *elliptically polarized*. From the vantage point shown by Fig. 7.14, the circular wave may rotate counterclockwise and is then called right-hand circular polarized light (apply the right-hand rule with the thumb pointing toward yourself), or it may rotate clockwise and is called left-hand polarized light.

This analysis makes it clear that a general harmonic wave propagating in the x direction has two components, one vibrating in the xy plane and one in the xz plane.

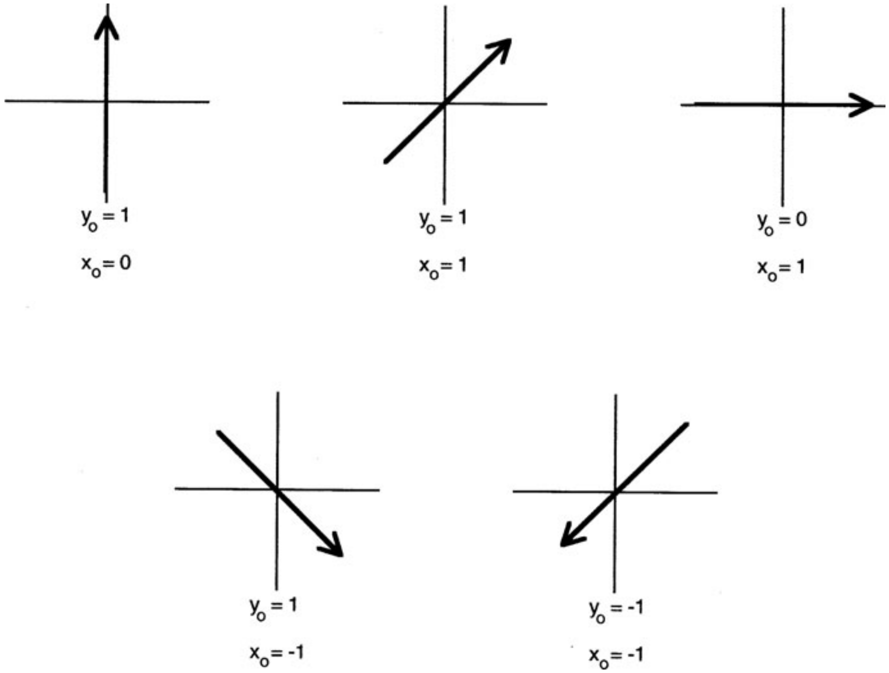


Fig. 7.13 Polarized waves. Varying the amplitudes of two plane polarized waves will rotate the resultant around the axis of propagation

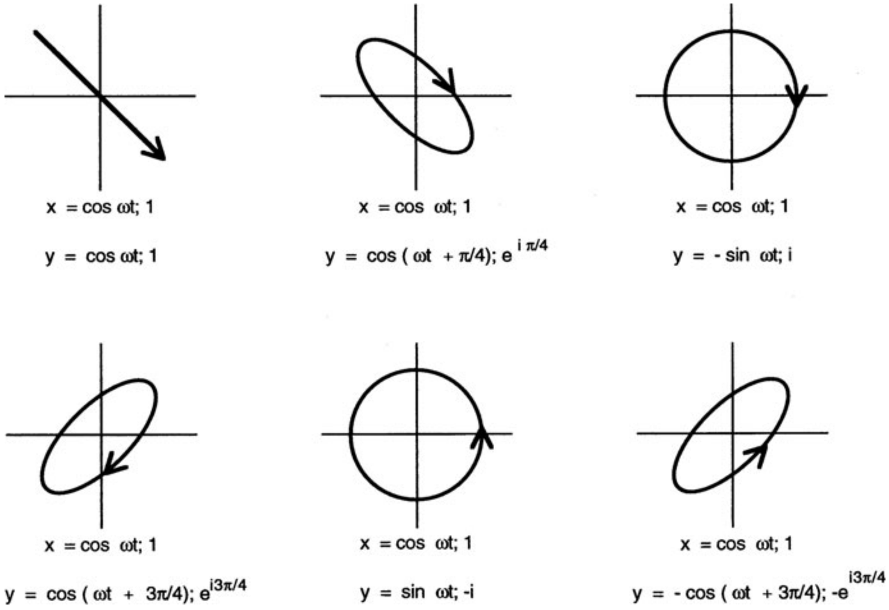


Fig. 7.14 Polarized waves. Varying the phase relation of two plane polarized waves will generate plane-, circularly, and elliptically polarized waves

Hence

$$\begin{aligned} y &= y_\circ \sin(kx - \omega t) \\ z &= z_\circ \sin(kx - \omega t + \delta) \end{aligned} \tag{7.105}$$

where δ is the phase difference between the two components. If the phase difference is constant then the wave is polarized. A wave is linearly polarized when δ is zero with the resultant wave falling along the line that makes an angle $\theta$with the z axis given by

$$\tan\theta = \frac{y_\circ}{z_\circ} \tag{7.106}$$

For the general case of elliptical polarization, of which circular polarization is just a special case, we can consider δ to be $\dfrac{\pi}{2}$ and

$$\begin{aligned} z &= z_\circ \sin\left(kx - \omega t + \frac{\pi}{2}\right) \\ &= z_\circ \cos(kx - \omega t) \end{aligned} \tag{7.107}$$

The y and z components can be related as

$$\begin{aligned} \frac{y^2}{y_\circ^2} + \frac{z^2}{z_\circ^2} &= 1 \\ &= \sin^2(kx - \omega t) + \cos^2(kx - \omega t) \end{aligned} \tag{7.108}$$

This is the general equation for an ellipse. When $y_\mathrm{o} = z_\mathrm{o}$ as in (7.107) the case of a circle is described.

The phase of a wave is frequently described on the imaginary plane as a matter of mathematical convenience and as a formalism. In Fig. 7.14 both the real and imaginary notations are used. The real x and y coordinates can be measured and have physical meaning while the imaginaries are formalisms that simply point to the observable quantities.

7.6.3 Superposition and Interference – Waves of the Same Frequency

The principle of superposition is fundamental in describing wave behavior in general and has broad application when discussing harmonic waves. The principle of superposition is the algebraic addition of two or more waves to give a resultant wave. Two harmonic waves traveling in the same direction with the same frequency and amplitude but differing in phase will give rise to a resultant harmonic wave of the same frequency and wavelength but with a new amplitude and phase that are dependent on the phase difference of the original waves. We can describe the original wavefunctions as follows:

$$\begin{aligned} y_1 &= y_\circ \sin(kx - \omega t) \\ y_2 &= y_\circ \sin(kx - \omega t + \delta) \end{aligned} \tag{7.109}$$

Superposition gives

$$\begin{aligned} y_3 &= y_1 + y_2 \\ &= y_\circ \sin(kx - \omega t) + y_\circ \sin(kx - \omega t + \delta) \end{aligned} \tag{7.110}$$

which can be simplified by using a trigonometric relation

$$\sin A + \sin B = 2 \sin \frac{1}{2}(A + B) \cos \frac{1}{2}(A - B) \tag{7.111}$$

this result, then writing $A = (kx - \omega t)$ and $B = (kx - \omega t + \delta)$ gives

$$\begin{aligned} &\tfrac{1}{2}(A + B) = \left(kx - \omega t + \tfrac{1}{2}\delta\right) \\ &\tfrac{1}{2}(A - B) = -\tfrac{1}{2}\delta \end{aligned} \tag{7.112}$$

and recognizing that $\cos\left(-\frac{1}{2}\delta\right) = \cos\left(\frac{1}{2}\delta\right)$, the resultant, y_3, can be written as

$$y_3 = 2y_\circ \cos \frac{1}{2}\delta \sin\left(kx - \omega t + \frac{1}{2}\delta\right) \tag{7.113}$$

The resultant is a new wave of identical frequency and wavelength with a new phase and an amplitude of $2y_{\text{o}}\cos \frac{1}{2}\,\delta$. When $\delta = 0$, $y_1 + y_2$ are in phase and $\cos \frac{1}{2}\,\delta = 1$, which means that the amplitude of the resultant wave is twice the amplitude of either of the original waves. This is the extreme of *constructive interference*. Alternatively if $\delta = \pi$, $y_1 + y_2$ are 180° out of phase and $\cos \frac{1}{2}\delta = 0$, which means that the original waves cancel each other completely (i.e., perfect *destructive interference)*.

How do we conveniently create two waves identical except for a phase difference? The answer is that we split a wave and force it down two paths. If each of the paths is different in length before being recombined, the rejoined waves will meet differing only by phase angle. If the difference in path length is exactly one wavelength, the phase difference is 2π and perfect constructive interference is seen (this condition is duplicated at any integer value times the wavelength). If the path difference is $\frac{1}{2}$ a wavelength, the phase difference will be π and complete destructive interference occurs (a condition which is duplicated at any odd integer value times the wavelength). The phase difference in general can be found from the difference in path length, Δx:

$$\delta = 2\pi \frac{\Delta x}{\lambda} \tag{7.114}$$

Vector addition of harmonic waves is a very convenient method for combining more than two waves and for considering waves of differing amplitudes. The vector addition methods are reviewed in Appendix A and are easily applied once a geometric interpretation of the wavefunction is considered. Wavefunctions are generally written either in real rectangular coordinate planes or in polar coordinate notation in which the amplitude of the wave is written in relation to one axis and the phase angle is described in relation to the other. These notations can often become cumbersome. Alternatively, it is often more convenient to write the amplitude of the wavefunction in a real plane and the phase angle in the imaginary plane. The latter is one of mathematical and notational convenience, and having phase angles in the imaginary plane has no physical meaning.
Example:
Find the resultant of two waves

$$y_1 = 3\sin(kx - \omega t)$$
$$y_2 = 6\sin(kx + \omega t + 70\Phi)$$

Solution:
See Fig. 7.15. Our task is to combine these waves at a particular point *x* and time *t*. The notation can be simplified by writing θ for kx – ωt. We can write

$$y_1 = 6\sin\theta$$
$$y_2 = 3\sin\left(\theta + 70^\circ\right)$$

and

$$\begin{aligned} y_3 &= y_1 + y_2 \\ &= 6\sin\theta + 3\sin\left(\theta + 90^\circ\right) \\ &= 6.708\sin\theta + 26.5^\circ \end{aligned}$$

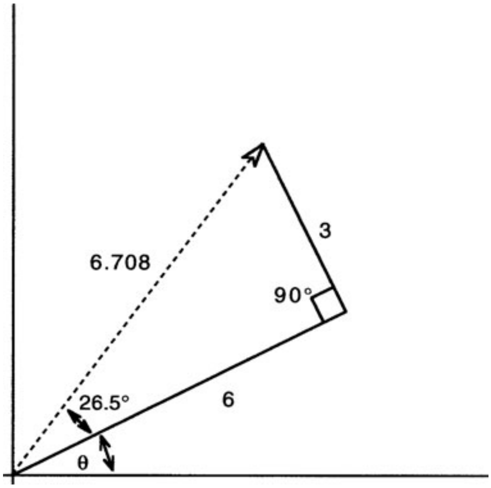


Fig. 7.15 Vector analysis of two waves to give a resultant wave

The angle θ is the constant frequency of the wave. It sweeps y_1, y_2, and y_3 through space together, and the relative positions of the wave vectors are not affected.

7.7 Energy and Intensity of Waves

Waves transport energy and momentum. The transport of energy in a wave is described by the wave intensity, I, which is defined as the average rate of transmitted energy per unit area normal to the direction of propagation. I is the average incident energy per unit time per unit area:

$$I = \frac{(\Delta E/\Delta T)_{\text{average}}}{A} = \frac{P_{\text{ave}}}{A} \tag{7.115}$$

where A is the cross-sectional area.

The intensity can be written in terms of power, P, since power is energy per unit time. How do we find ΔE? It will be remembered that the energy in the wave is being supplied by an external force that is causing the wave motion. Consider this energy source to be inducing a harmonic wave motion in a string at the left end. If we look down the string and observe a point in front of us before the wave reaches that point, it is obvious that the string to our left contains energy but the string to our right does not yet contain this energy. As the wave sweeps past us at a velocity, v, in a certain time, Δt, it travels a distance equal to $v\Delta t$. There has now been an increase in energy carried in the wave, ΔE, which is equal to the average energy per unit volume of the string, η, multiplied times the unit volume of the string which now contains energy, namely

$$\Delta E = \eta A v \Delta t \tag{7.116}$$

The rate of energy increase, $\Delta E/\Delta t$, is the power. Combining Eqs. (7.115) and (7.116), the average incident power can be stated as

$$P_{\text{ave}} = \frac{\Delta E}{\Delta t} = \eta A v \tag{7.117}$$

The intensity of the wave at a point is the incident power divided by the area of the string (or propagating medium):

$$I = \frac{P_{\text{ave}}}{A} = \eta v \tag{7.118}$$

Equation (7.118) is a general result applicable to all waves.

What is the average energy density, η? We seek a general answer that is applicable to all harmonic waves. For a harmonic wave on a string with angular frequency ω and amplitude y_{o}, each mass element on the string moves with the energy of a mass oscillating on a spring (see earlier discussion):

$$\Delta E = \frac{1}{2}\,(\Delta m)\,\omega^2 y_\circ^2 \tag{7.119}$$

The mass of the segment can be written in terms of the mass density times volume of the segment, $\rho\Delta V$:

$$\Delta E = \frac{1}{2}\,(\rho\Delta V)\,\omega^2 y_\circ^2 \tag{7.120}$$

and substituting

$$\eta = \frac{\Delta E}{\Delta V} = \frac{1}{2}\rho\omega^2 y_\circ^2 \tag{7.121}$$

so

$$I = \eta v = \frac{1}{2}\rho\omega^2 y_\circ^2 v \tag{7.122}$$

This general result states that the intensity of a wave is proportional to the square of its amplitude and the square of its frequency. A direct application of Eq. (7.122) is in considering the intensity of sound, which can be a significant issue in biological systems whether in using ultrasound to enhance transdermal diffusion of pharmaceuticals or in considerations of noise pollution.

7.7.1 Sound and Human Ear

The human ear is able to respond to a large range of sound intensities from approximately 10^{-12} W/m^2 to about 1 W/m^2. The lower end corresponds to the threshold of hearing while sound with an intensity of 1 W/m^2 causes pain. The ear and the brain respond to sound intensity by the sensation of loudness. Loudness varies in a logarithmic relationship to the sound intensity. Hence given the large magnitude of the range of hearing as well as the psychological interpretation of intensity as loudness, a logarithmic scale, the decibel scale, is used to describe the intensity of a sound wave. Intensity is measured in *decibels*, dB, which correspond to the intensity β in the following way:

$$\beta = 10\log\frac{I}{I_\circ} \tag{7.123}$$

I is the intensity which is described in dB and it is related to the threshold of hearing (I_o) which is taken to be 10^{-12} W/m^2. Using the decibel scale, the threshold for hearing is 0 dB and the threshold for pain is 120 dB. Sound intensities for a variety of environmental sounds are often surprising when compared on the decibel scale (see Table 7.6).

Table 7.6 Intensity of common sounds

Sound	dB
Normal breathing	10
Soft whisper at 5 m	30
Quiet office	50
Busy traffic on street	70
Heavy truck at 15 m	90
Pneumatic jack hammer	130
Hearing threshold	0
Rustling of leaves	20
Quiet library room	40
Normal voice conversation	60
Factory/busy office	80
Subway train	100
Jet takeoff at runway	150

7.8 Standing Waves

If waves are confined within a specific space, the waves will propagate in one direction and then be reflected back onto themselves. If the wavelength and constrained space are correctly chosen, the reflected waves and the propagating waves will constructively interfere and give a special pattern of interference called *standing waves*. It is standing waves that make the sounds of a pipe organ, violin, or flute. When a propagating wave meets its reflection the resulting interference leads to an ever increasing amplitude of the standing wave. This continues until the imperfections in the system cause a damping of the progressive increase at a maximum amplitude. At this maximum, a state of *resonance,* between the exciting agent (such as a tuning fork) and the system in which the standing wave is formed (the string or pipe), is said to have occurred. Resonance occurs when a wave propagated in a string travels the length of the string, is inverted and reflected and strikes the propagating end (in this case attached to a tuning fork), and is again reflected. The twice reflected wave is inverted again and so is upright when compared to the new wave now being propagated by the tuning fork. The newly propagated wave and the doubly reflected wave are of the same phase and differ only in that the first wave has already traveled twice a distance that is twice the length of the string or $2L$. The resonance condition is thus achieved whenever twice the distance traveled is equal to the wavelength or any integer multiple of the wavelength:

$$2L = n\lambda \tag{7.124}$$

It is valuable at times to consider the lowest possible frequency that can lead to a resonance condition; this frequency, f_1, is called the *fundamental frequency*. From $\lambda = \frac{v}{f}$ we can write for any frequency that will cause resonance:

$$f = n\frac{v}{2L} = nf_1 \tag{7.125}$$

where f_1 can be found if the velocity of the wave is known. Using the earlier result from Eq. (7.96) we can write

$$f_1 = \frac{v}{2L} = \frac{1}{2L}\sqrt{\frac{T}{\mu}} \tag{7.126}$$

The integer multiples of the fundamental frequency (given by (7.125)) are called the *natural frequencies*.

The physical consequence of the resonance condition is that maximum energy can be absorbed from the active element such as the tuning fork. If the frequency of the tuning fork does not equal one of the natural frequencies of the string, then the wave generated by the tuning fork initially will generally remain at the amplitude of the fork. The reason is that over a period of time, the interference of the reflected and newly propagated waves will add to zero and *on average* no energy will be absorbed by the string. You can convince yourself of the truth of this statement by graphing this problem and roughly calculating the resultant waves.

If a string is fixed at both ends, there is not just one natural frequency but a series of natural frequencies which are integral multiples of the fundamental frequency. This sequence is called a *harmonic series*. In a state of resonance, the driving element transfers energy at a natural frequency to the string such that the waves add constructively up to a maximum, which is ultimately limited by the damping forces present in the system. A wave is produced on the string that appears to stand still and is called a standing wave. The standing wave shapes for a portion of a harmonic series are shown in Fig. 7.16. In a standing wave system, the energy of the wave is constant. The energy varies between a maximum potential energy when the wave causes the string to be maximally displaced, and a maximum kinetic energy when the string is at the original coordinate but each segment of the string is moving at maximum speed.

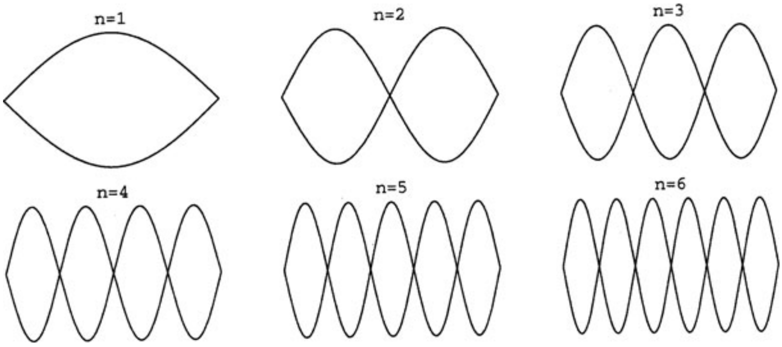


Fig. 7.16 This harmonic series shows the nodes and antinodes of the standing waves generated as multiples of the fundamental frequency where n=1

The mathematical form of standing waves can be found by simply finding the resultant wave derived from a wave propagated from the left, y_{L}, and one propagated from the right, y_{R}. For simplicity we will let the amplitudes be equal and we can write

$$\begin{aligned} y_R &= y_\circ \sin\left(kx - \omega t\right) \\ y_L &= y_\circ \sin\left(kx + \omega t\right) \end{aligned} \tag{7.127}$$

The sum of the waves is

$$\begin{aligned} y\left(x, t\right) &= y_R + y_L = y_\circ \sin\left(kx - \omega t\right) + y_\circ \sin\left(kx + \omega t\right) \\ &= 2y_\circ \cos\left(\omega t\right) + \sin\left(kx\right) \end{aligned} \tag{7.128}$$

Equation (7.128) is the function for a standing wave. We can show this equation to give the resonance condition if we fix the string at $x = 0$ and $x = L$ thus forcing y at these points to be zero for all times t. We will use a similar treatment when we determine the probability wave for a particle captured in a box. The boundary conditions can then be written out:

$$\begin{aligned} y\left(x = 0, t\right) &= 0 \\ y\left(x = L, t\right) &= 0 \end{aligned} \tag{7.129}$$

Since at $x = 0$, sin $kx = 0$, the first condition is automatically met. However, there are only a selected set of values of the wave number, k, for which the second boundary condition, sin $kL = 0$, will be satisfied. Since the sin of a number is only zero at sin 0 or at the sin of an integer multiple of π radians, the values of k which will satisfy the second condition are given by

$$kL = n\pi \tag{7.130}$$

Writing this in terms of wavelength where $k = \dfrac{2\pi}{\lambda}$,

$$\frac{2\pi}{\lambda} L = n\pi \tag{7.131}$$

gives the same result as the resonance condition equation (7.124):

$$2L = n\lambda \tag{7.132}$$

Since the standing wavefunctions are sine waves that must fit into the boundary conditions when $n > 1$, there will be other points that are at rest in addition to the two endpoints. These points are called *nodes*, and there are generally $n - 1$ nodes in addition to the boundary points at rest for the nth harmonic. The points of maximal displacement are called *antinodes* and there are n antinodes for the nth harmonic.

7.9 Superposition and Interference – Waves of Different Frequencies

We have thus far considered only the cases of superposition of waves of equal frequency. Waves of differing frequency will also generate resultant waves, but the resultant will often result in a complex waveform. The detailed analysis of complex waveforms uses the mathematical techniques of Fourier analysis which are beyond the scope of this volume. Here we will be content with a qualitative review of the subject.

The treatment of the superposition of different frequencies is similar to the analysis already undertaken. Consider the following case which we will express both graphically and algebraically. Two waves of equal amplitude and both moving left to right possess different frequencies. We will consider the case of frequencies close to one another. Initially, (i.e. when $x = 0$ and $t = 0$) both waves are in phase. Initially the waves will add to one another, constructively giving rise to a large resultant amplitude, but since they are out of phase, at some distance x_1, their phase difference will be 180°, and there they will interfere destructively with a resultant amplitude of zero. At an equal distance x beyond x_1, at x_2, the waves will again be in phase with maximal constructive interference. The greater the difference in wavelength between the two waves, the shorter will be the distance x. Algebraically we can write

$$\begin{aligned} y_1 &= y_\circ \cos\left(k_1 x - \omega_1 t\right) \\ y_2 &= y_\circ \cos\left(k_2 x - \omega_2 t\right) \end{aligned} \tag{7.133}$$

Using the cosine identity, $\cos A + \cos B = 2\cos\frac{1}{2}(A-B)\cos\frac{1}{2}(A+B)$ the resultant wave, $y(x, t)$ can be written as

$$\begin{aligned} y\left(x,t\right) &= y_2 + y_1 \\ &= y_\circ \cos\left(k_2 x + \omega_2 t\right) + y_\circ \cos\left(k_1 x - \omega_1 t\right) \\ &= 2y_\circ \cos\left[\frac{1}{2}\left(k_2 - k_1\right)x - \frac{1}{2}\left(\omega_2 - \omega_1\right)t\right] \\ &\quad \cos\left[\frac{1}{2}\left(k_2 + k_1\right)x - \frac{1}{2}\left(\omega_2 + \omega_1\right)t\right] \end{aligned} \tag{7.134}$$

Now we substitute some new notation to clean up the equation, using Δk and $\Delta\omega$ for the difference in wave number and frequency and $\overline{k}$ and $\overline{\omega}$ for their averages, respectively.

$$\Delta k = \left(k_2 - k_1\right) \;\; \Delta\omega = \left(\omega_2 - \omega_1\right)$$

$$\overline{k} = \frac{1}{2}\left(k_2 + k_1\right) \;\; \overline{\omega} = \frac{1}{2}\left(\omega_2 + \omega_1\right)$$

which allows writing

$$y\,(x,t) = 2y_{\circ}\cos\left[\frac{1}{2}\Delta kx - \frac{1}{2}\Delta\omega t\right]\cos\left[\overline{k}x - \overline{\omega}t\right] \qquad (7.135)$$

Now writing

$$y(x,t) = 2\mathrm{y}_{\mathrm{o}}\cos\,[\frac{1}{2}\Delta kx - \frac{1}{2}\Delta\omega t]\cos\,[\hat{k}\,x \text{ - } \omega]t \qquad (6.57)$$

This resultant wave has a frequency about the same as the original waves but modulated by a factor, $\cos\left[\frac{1}{2}\Delta kx - \frac{1}{2}\Delta\omega t\right]$.

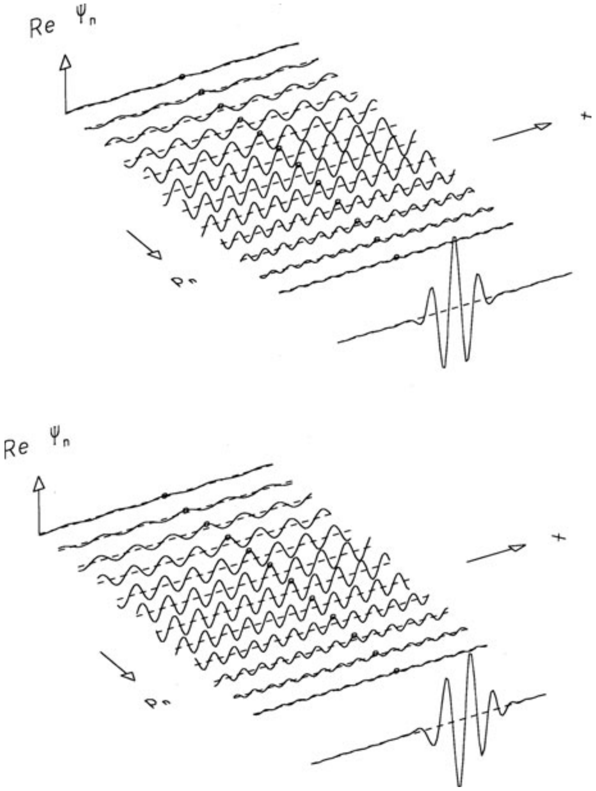


Fig. 7.17 Waves of different frequency interfere in a complex way and are characterized by a phase and group velocity. Here, a Gaussian distribution of waves is combined to give the resultant wave packets. The top figure shows the wave packet at $t = 0$ and the lower figure shows the same wave packet at a time Δt later. Notice that the wave packet has moved along the x-axis in a fashion reminiscent of a particle. However, this illustration also shows some moderate dispersion of the wave packet, and the different phase and group velocity are causing the wave packet to change shape with time

The velocity of the resultant wave is

$$v_p = \frac{\overline{\omega}}{k} \tag{7.136}$$

and is called the *phase velocity*. The phase velocity is nearly the same as the initial waves. The envelope of modulation outlined in Fig. 7.17 travels as a wave itself with a wave number equal to $\frac{1}{2}\Delta k$ and angular frequency of $\frac{1}{2}\Delta\omega$. The envelope can be shown to have the form of a wavefunction by considering the modulation factor:

$$\cos[\frac{1}{2}\Delta kx - \frac{1}{2}\Delta\omega t] = \cos\frac{1}{2}\Delta k\left(x - \frac{\Delta\omega}{\Delta k}t\right) \tag{7.137}$$

and defining the *group velocity*, $v_g = \dfrac{\Delta\omega}{\Delta k}$. Equation (7.137) can be written as

$$\cos\frac{1}{2}\Delta k\left(x - v_g t\right) \tag{7.138}$$

This equation has the mathematical character of *a* and hence the group moves through space as if it were a wave itself. The group velocity and phase velocity are equal when the medium through which the wave is propagated is *dispersionless*, i.e., the phase velocity does not depend on the frequency of the wave. Light waves in a vacuum, sound in the air, and waves in a perfectly flexible string will not disperse. A medium in which the phase velocity depends on frequency is called a *dispersive medium*, and light traveling in glass or water waves are examples. As we will see below, this group as represented by the enveloped waves forms a distinct wave packet with important applications in information theory, cybernetics, and quantum mechanics (see Fig. 7.17).

7.10 Complex Waveforms

Complex waves are in general well treated as mixtures of harmonic waves of varying frequencies. Fourier's theorem states that any periodic function can be described to an arbitrary degree of accuracy in terms of the sum of sine and cosine functions:

$$y(t) = \sum_n A_n \sin\omega_n t + B_n \cos\omega_n t \tag{7.139}$$

The lowest angular frequency in the analysis will correspond to the period of the function under study, $\omega_1 = \frac{2\pi}{T}$, where T is the period. All the other frequencies will be integral multiples of this fundamental frequency. The values of A_n and B_n represent the relative intensity of each of the harmonics found in the waveform and are found by Fourier analysis. The value of $A_n^2 + B_n^2$ is proportional to the energy of the nth harmonic of the function. The intensity of each harmonic present can be plotted against frequency of each harmonic wave present to provide a spectral

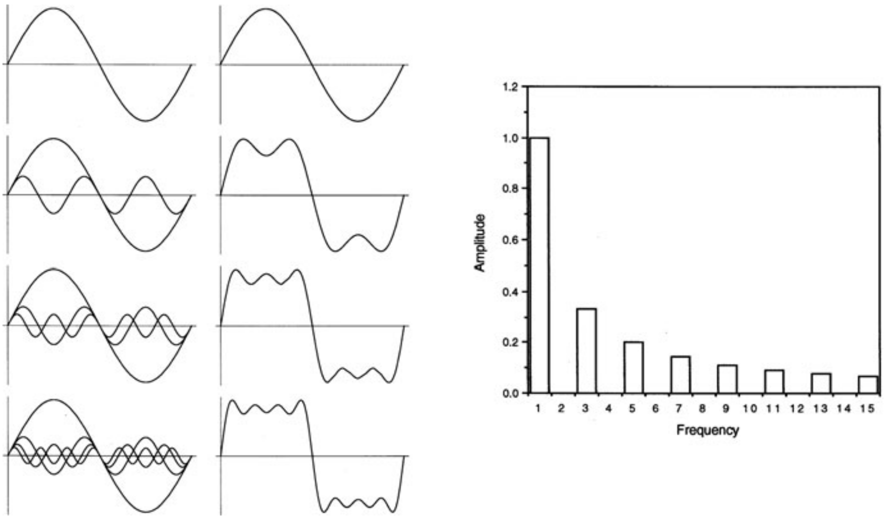


Fig. 7.18 An example of a Fourier analysis. A *square wave* is a converging geometric sum of the amplitudes of the odd harmonics of the fundamental frequency. (**a**) The *left panels* show the harmonic waves that create the square wave of a specific duration. The *right panel* shows the superposition of those waves. Note how increasing the number of harmonics summed together improves the Fourier representation. (**b**) This figure shows a power spectrum that relates the frequencies and intensities of each of those frequencies contributing to the square wave

analysis of the constituent waves in a complex function. Plots of this type are called a *power spectrum*. An example of the harmonic analysis and power spectrum of a square wave pulse is shown in Fig. 7.18.

7.11 Wave Packets

Let us now revisit the wave impulses with which we started our review. In the last section we were using harmonic analysis to examine complex waveforms that were none-the-less periodic in time. Non-periodic functions (i.e., pulses) can also be described in terms of sine and cosine functions, but to do so requires a continuous distribution of frequencies rather than the discrete summation of the Fourier theorem. Mathematically we must integrate over frequency and we will find that the constants A_n and B_n become functions of frequency, $A_{(\omega)}$ and $B_{(\omega)}$. These constants are often close to zero over most of the frequency range but will have non-zero value over a specific range $\Delta\omega$. Pulses are quite distinct from harmonic waves in that they clearly begin and end while harmonic waves continue repeating over and over. It is the starting and stopping of a pulse-like signal that allows information to be sent by a wave. The relationship between distribution of the harmonic frequencies that make up a pulse and the duration of the pulse is important. If the duration of a pulse is very short, then the range of frequencies $\Delta\omega$ must be very large. The relationship between frequency and time can be written as follows:

$$\Delta t \Delta\omega \approx 1 \text{ and } \Delta\omega \approx \frac{1}{\Delta t} \tag{7.140}$$

A wave pulse of duration Δt occupies a narrow band of space determined by the product of the time and the velocity of the wave,

$$\Delta x = v\Delta t \tag{7.141}$$

The harmonic waves in the interval $\Delta\omega$ that comprise the pulse each have a wave number: k,

$$k = \frac{\omega}{v} \tag{7.142}$$

and thus in $\Delta\omega$, there exist a range of wave numbers:

$$\Delta\omega = \Delta k v \tag{7.143}$$

thus we can write

$$\Delta k v \Delta t \approx 1 \text{ or } \Delta k \Delta x \approx 1 \tag{7.144}$$

The relationships in Eqs. (7.139) and (7.144) are important in communications theory and quantum mechanics, both topics of importance in understanding biological systems at a molecular chemical and molecular physiological level. These relationships are very important both practically and philosophically since they relate quantities that are called *complementary*. The relationship between energy and time is complementary as are the observables position and momentum. The former is seen in Eq. (7.140) where ω is proportional to the energy of the packet and the latter in Eq. (7.144) where the wave number is proportional to the momentum of the packet. The concept of complementary observables is fundamental in understanding the constraints on what is knowable in a physical system: Complete knowledge of one aspect of the system makes any knowledge of the complementary observable impossible.

In communications theory, information cannot be carried by an unvarying continuous harmonic wave. Maximal information is encoded by short pulses. The shorter the pulse, the wider the bandwidth of the transmitting system and, hence the larger the amount of information that can be transmitted in a fixed time. Maximizing information transfer requires the shortest possible pulse, but this means that the short pulse contains a very large number of harmonics. A system with the capacity to handle this wide range of frequencies without dispersion or distortion must be used. This is a primary concern in engineering communication systems both man-made and biological.

In quantum mechanics the impossibility of knowing values for a pair of complementary observables such as position and momentum is the core of the Heisenberg indeterminacy principle. The location in space of a particle can be described in terms of a wave packet whose width is interpreted as the uncertainty in knowing the position of the particle. A very sharply defined and narrow wave packet, which leads to a high degree of certainty in the position of the particle, must at the same time possess a large number of associated wave numbers which are representative of the momentum of the particle. The narrower the wave packet, the more certain we are of the position. However, this condition also results in a greater number of possible wave numbers, which represent the momentum values the particle may take, and so we are less certain of the particle's momentum.

7.12 Dispersion

The envelope of two waves of unequal frequency discussed above is a simplified case of more complex waveforms. We will now examine qualitatively the important and interesting effects of wave motion in a medium that is dispersive. A waveform that travels in a dispersionless medium has a phase velocity and group velocity that are equal. The equations written as (7.136) and (7.138) were derived for two waves, and the case can be generalized to include complex waves and wave packets. In a wave packet, a large number of harmonic waves are superimposed over a limited range of wavenumbers, k, and angular frequencies, ω. Within the interval of interest, k and ω are functions dependent on one another, and we can write that ω is a function of k. In order to write an equation that is more general than the case of just two waves we will evaluate the function in the midpoint of the range of k or ω:

$$v_g = \frac{\Delta\omega}{\Delta k} \approx \frac{dw}{dk} \tag{7.145}$$

This is the generalized equation for the group velocity. What is the relationship between each constitutive harmonic wave and the overall wave packet? Since each harmonic has its own wave number and angular frequency and travels at a phase velocity, we can write

$$v_p = \frac{\omega}{k} = f\lambda \tag{7.146}$$

Now we use this relationship to find the relationship between the group velocity and the phase velocity of each constitutive harmonic:

$$v_g = \frac{dw}{dk} = \frac{d}{dk}\left(kv_p\right) \approx v_p + k\frac{dv_p}{dk} \tag{7.147}$$

A wave packet will change shape if the individual harmonic waves travel at speeds different from the speed of the overall packet. When these shape changes occur it is because the velocity of the wave is dependent in some fashion on the wave number or its related quantities, the wavelength, or frequency. Light waves traveling through a vacuum do not disperse because there is no interaction along the way; but a light wave in any transparent medium does interact in a frequency-dependent manner and will be found to be dispersed by the transparent medium. A dispersion equation can be written that links the wavenumber of a light wave to its speed in transparent media, and this dispersion relationship defines the refractive index of the medium. The dispersion relationship can be understood in terms of the interaction of the wave with the elements of the medium.

7.13 The Wave Equation

The wave equation recurs throughout our studies of biophysics and some degree of familiarity with it is important. We will discuss it here as a one-dimensional problem.

The wave equation is a differential equation that results directly from Newton's second law; it relates the second derivative of a wavefunction with respect to x to the second derivative with respect to time. Note that because there are two variables the wave equation will be the result of a partial differential equation. The general wavefunction *y(x,t)* is a solution of this equation. For waves on a string or in a fluid the wave equation can be derived from the second law with the restriction that the wave is of small enough amplitude so that the small angle approximation can be made. Therefore the equations remain linear. For electromagnetic waves, the wave equation can be shown to derive directly from the laws of electricity without the small amplitude approximation.

We will show the wave equation as a result of the harmonic wavefunction.

We start with the harmonic wavefunction given in Eq. (7.97):

$$y(x,t) = y_\circ \sin k(x - vt) \tag{7.97a}$$

We multiply through by k and take the derivative with respect to x while holding t constant:

$$\frac{\partial y}{\partial x} = ky_\circ \cos\,(kx - kvt) \tag{7.148}$$

and differentiating again

$$\frac{\partial^2 y}{\partial x^2} = -k^2 y_\circ \sin\,(kx - kvt) \tag{7.149}$$

Now look at this result carefully! We have differentiated a function twice and *gotten the same function back* times a constant. This is an important property of the class of equations we saw in the previous chapter, *eigenfunctions*. We can now write Eq. (7.149) in terms of the original function:

$$\frac{\partial^2 y}{\partial x^2} = -k^2 y\,(x,t) \tag{7.150}$$

Now we differentiate the other partial derivative, $y(x,t)$ with respect to t with x held constant:

$$\frac{\partial y}{\partial t} = -kvy_\circ \cos\,(kx - kvt) \tag{7.151}$$

and once again for the second derivative

$$\begin{aligned}\frac{\partial^2 y}{\partial t^2} &= -k^2 v^2 y_\circ \sin\,(kx - kvt)\\ &= -k^2 v^2 y\,(x,t)\end{aligned} \tag{7.152}$$

Another eigenfunction! These can be combined to form the wave equation:

$$\frac{\partial^2 y}{\partial x^2} = \frac{1}{v^2}\frac{\partial^2 y}{\partial t^2} \tag{7.153}$$

For a wave on a string y represents the displacement of the string, for a sound wave y is the pressure or density change, and for an electromagnetic wave y is the electric or magnetic field magnitude. Equation (7.82) applies to any wave propagating in one dimension without dispersion that can be shown by evaluating the more general functions given in Eqs. (7.141) and (7.143).

The wave equation is linear (i.e., it has no roots or derivatives that are to any power except the first power). The principle of superposition naturally derives from this linearity since linear equations have the property that if $y_1\,(x,t)$ and $y_2\,(x,t)$ are solutions of the equation, the linear combination

$$y\,(x,t) = C_1 y_1\,(x,t) + C_2 y_2\,(x,t) \tag{7.154}$$

is also a solution. Thus any two waves satisfying the wave equation can be simply added and their resultant will also satisfy the equation.

7.14 Waves in Two and Three Dimensions

Now we generalize our discussions of one-dimensional waves to two and three dimensions which are the wave systems with which we more commonly deal. Consider a point source generating a harmonic wave. Figure 7.19 shows how a circular (two-dimensional) or a spherical (three-dimensional) wave is generated with a point source as the origin. In the case of a spherical wave, the first sphere at distance r from the origin has an area of $4\pi r^2$. The energy is evenly distributed

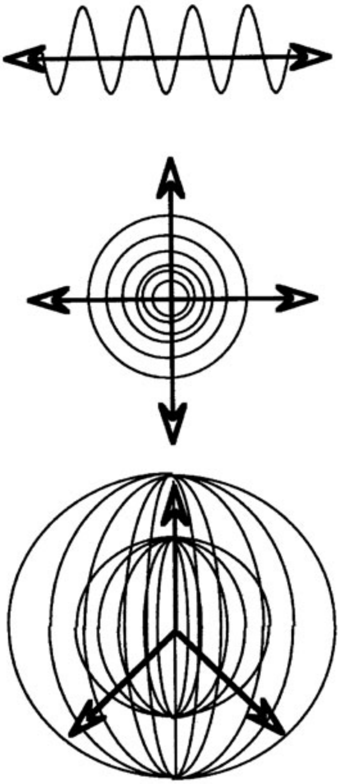


Fig. 7.19 The shape of a propagating wave in (**a**) one, (**b**) two, andg (**c**) three dimensions

over this area. At each subsequent sphere *nr* from the origin, the same energy is distributed over an area of $4\pi nr^2$ so the intensity decreases as $\frac{1}{r^2}$. This relationship holds even when energy is not equally radiated as can be shown if a cone of energy is considered.

Since the intensity of a wave is proportional to the square of its amplitude, the amplitude of a spherical wave must vary as $\frac{1}{r}$. The wavefunction, ψ, of a spherical wave is written as

$$\psi(r,t) = \frac{A_\circ}{r}\sin(kr - \omega t + \delta) \qquad (7.155)$$

where the constant A_o is independent of r, and δ, as before, is a phase constant. The phase of this wave is constant on a spherical surface at any given time. Such a surface of constant phase is called a *wavefront*. Wavefronts are usually drawn at the maximum amplitude of the wave, and the radial distance between successive maxima is the wavelength. The phase velocity v is calculated as $v = f\lambda$. A very convenient rendering of spherical waves is the ray tracing (Fig. 7.20a), and we can

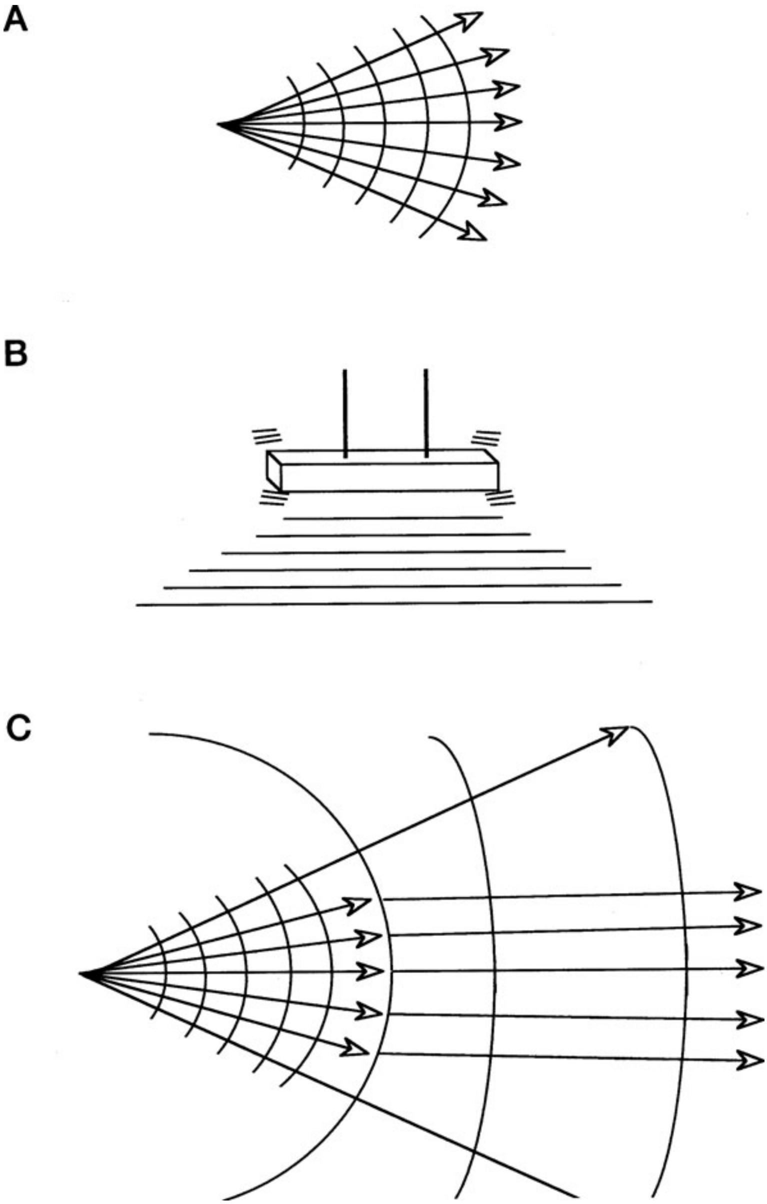


Fig. 7.20 Representation of (**a**) the ray tracing abstraction of a spherical wave. (**b**) Ripple tank production of a plane wave. (**c**) At a distance from the point of propagation, a spherical wave can be treated as a plane wave

see both the advantages and limitations of such an *abstraction* in Appendix G. When the wavefronts are far away from the point source, the curvature of the wavefront is so small that the ray is essentially perpendicular to the wavefront, and the spherical wave appears to be a *plane wave* (Fig. 7.20c). A two-dimensional representation of a plane wave can be generated in a ripple tank by the oscillation of a flat object up and down (Fig. 7.20b). When a spherical wave has reached the point at which it can be regarded as a plane wave, the wavefronts move perpendicular to the *x* axis in a series of *yz* planes. At this point the phase depends only on *x* and *t* and the wavefunction has the same form as a one-dimensional wave:

$$\psi(x,t) = A_{\circ} \sin(kx - \omega t + \delta) \tag{7.156}$$

Plane waves traveling in a featureless and unobstructed medium will act in identical fashion to the one-dimensional waves that we have examined. If they strike an infinitely long boundary in the *yz* plane, the wave can be treated as being reflected like a one-dimensional wave. However, if a plane wave runs into apertures or obstructions, the plane symmetry is destroyed and it will act once again like a three-dimensional wave. The consequences of these distortions lead to reflection, refraction and diffraction and are phenomena with immensely important implications both theoretically and practically.

Further Reading

See the listed readings in Chapter 6.

Problem Sets

1. (a) Calculate the proportionality coefficient of the ideal gas law, *R*, the gas constant from the ideal gas law using STP. (b) Use the conversion factors listed in the appendix to express *R* in units of J/K/mol.
2. Use the van der Waals equation to determine the molecular size of a gas atom of helium, neon, and water vapor.
3. Ion channels allow ion flow through the lipid membrane of a cell. Many channels are selective allowing only certain ions through. Some channels have a selectivity filter that prevents anions even of the appropriate charge from penetrating into a channel reserved for cationic species. The selectivity filter can be modeled as a ring of charge around the mouth of the channel. Calculate the force on a point charge as it moves on the axis of a uniformly distributed ring of charge. Calculate for an anion of charge –1, −2 and a cation of charge +1, +2.
4. Calculate the force due to a spherically symmetrical charge distribution, i.e., a cell interacting with an ion or proton.
5. Calculate the electric field at distance *R* from an infinite sheet of charge with uniform surface charge.

6. A popular act of violence in the movies is the electrocution of a character while they are taking a bath by having an electrical appliance such as a hair dryer or lamp thrown into the tub of water.

   a) Draw a potential map showing the isoelectric lines for this situation.
   b) Is it fact or fantasy that the act of violence is successful?

Chapter 8
Physical Principles: Quantum Mechanics

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_8, © Springer Science+Business Media, LLC 2010

8.1 The Story of the Discovery of Quantum Mechanics Is an Instructive History of How Scientific Ideas Are Modified

The practitioners of Western physics believed that they had reached the climax of their art by the end of the nineteenth century. When a young Max Planck turned to the study of physics, he was warned that the field was virtually closed and that no significant or fundamental discoveries remained to be made. These views are certainly understandable taken from the perspective of the nineteenth-century practitioner. Appreciating the certainty of this view and then following the path by which the ideas of classical physics were replaced provides one of the best historical examples confirming the ideas put forth in the first section. The quite surprising rise of modern physics (which is comprised of quantum and relativistic theory) from the monolith of classical physics followed from a fundamental shift in assumptions and a fresh view of the physical state space. As we will see in the following discussion, a shift in the paradigm from the macroscopic to the microscopic nature of the Universe leads to a reevaluation of the empirical evidence already at hand. The experimental evidence could not be understood (or better, properly linked to a formal model) until the simplifying assumptions on which the treatment of mechanics was based were revisited and revised. The new mechanical constructs that reflect both a shift in the way the state space is viewed and how the observables are linked to a formal theoretical model compose the language and concepts of modern physics. As such they are the foundation of modern natural science. We will follow the historical development of the quantum theory as an example of model making and will use this perspective as our approach for learning the details of the field.

8.2 From the Standpoint of the Philosophy of Epistemological Science, the Quantum Revolution Ended an Age of Certainty

Building on the foundations laid in the seventeenth and eighteenth centuries, classical physics came of age rapidly during the nineteenth century. Inspection of the time line in Table 8.1 provides an overview of the velocity of these developments.

Table 8.1 Time line of selected advances in physics over the period from 400 BCE to 1748 CE. The accelerating pace of events continues from 1750 through to the present time

400 BCE	Democritus proposes objects in the external world radiate beams that induce perceptions in the mind
350 BCE	Aristotle and Strato note acceleration of falling bodies. Heavy bodies are thought to fall faster
260 BCE	Archimedes discovers specific gravity
250 BCE	In the *Mo Ching*, the principle of inertia is proposed
100	The Chinese invent the compass
140	Ptolemy introduces epicycles to explain errors in the observed motion of planets
1110	Abu’l Futh al-chuzin writes tables of specific gravity
1270	Bacon studies optics and refraction by a lens
1490	Da Vinci describes capillary action
1543	Copernicus proposes a heliocentric solar system
1580	Stevinus studies hydrostatics and corrects Aristotle’s notion of the motion of falling bodies
1592	Galileo invents an early thermometer
1600	Kepler begins his work
1624	Helmont coined the word “gas” [Flemish for *chaos*] He discovers CO_2
1638	Galileo writes on the laws of motion and friction
1640	Torricelli applies Galileo’s laws to fluids and invents hydrodynamics
1666	Newton studies gravitation, color, light
1673	Huygens publishes on pendulums, centripetal force, and the conservation of kinetic energy
1675	Oemer makes an estimate of the speed of light
1676	Hooke’s law
1678	Huygens proposes light consists of waves
1704	Newton contends light is particulate
1705	Experimenting with a vacuum, Hauksbee shows sound requires a medium to be transmitted
1706	Hauksbee produces static electricity
1723	*Prodomus crystallographiae* by Capeller is published
1729	Gray proposes some materials to either conductors or insulators of electricity
1733	du Fay discovers that like charges repel, notes similarities to magnetism, proposes two separate fluids
1739	Martine shows heat content is not proportional to volume
1744	de Maupertuis formulates the principle of least action
1747	Franklin disputes du Fay, argues that relative excess and deficiency of electrical fluid leads to observations
1747	d’Alembert studies vibrating strings and publishes the solution to partial differential wave equations in two dimensions
1748	Nollet discovers osmosis
1748	J.O. de La Mettrie publishes *L’homme-machine* (*Man as Machine*). He argues that body and soul are mortal. Life and thought are nothing more than the mechanical action of the nervous system

Utilizing the natural philosophical investigations and mathematical work of the earlier scientists the nineteenth century physicists codified the laws that governed the large-scale aspects of matter and energy.

The almost completed edifice of classical physics was built on the success of theoretical laws and their practical value in describing the considered Universe. Using Newton's laws of gravitation and mechanics, the movements of the planets and stars could be predicted with confidence. Astronomers were able to predict and find the outer planets of Uranus and Neptune. The ability of the laws of thermodynamics to demonstrate the relationship between energy, work, and heat was practically realized in the steam and internal combustion engines. Electricity and magnetism were being explained and exploited through the treatment of continuum fields and wave propagation of the energy. Great electrical industries were built. Maxwell formulated his wave equations in 1865 to show that a changing electrical field will generate a magnetic field and vice versa. He calculated that the speed of electromagnetic radiation matched the speed of light. This led him to propose that light is electromagnetic in nature. This was the first of the theoretical unifications of physical phenomena.

For all of the majestic power, grace, and usefulness of classical physics, its perspective was largely macroscopic, and it was assumed that the structure of matter was straightforward and simple. Matter was made of indestructible, irreducible atoms that were eternal. In 1869, Mendelev had elucidated the periodic patterns of the physical properties of atoms when arranged by their atomic weight. Still this area of study was largely ignored by physics at the time.

The fundamental view of the state space by classical physics can be summarized succinctly by the terms *determinism* and *continuum*. In any classical system, once the initial positions, momentums, and directions of movement were described, *all* subsequent positions and interactions could be predicted with certainty simply by the appropriate application of the laws of mechanics, thermodynamics, and electromagnetics. A fundamental assumption was that the quantities for position, momentum, and energy taken by these systems were on a continuum. Therefore motion and position and, consequently, all interactions could be described by continuously differentiable functions with an infinite variety of smoothly assigned values.

8.3 The Ultraviolet Catastrophe Is a Term That Refers to a Historical Failure of Classical Theory

In spite of its successes, by the end of the nineteenth century classical physics was in disarray because of a variety of observations. Although most of the problems were in the areas of atomic and molecular phenomena, large-scale astrophysical phenomena were becoming problematic as well. The existence of the ether in which all electromagnetic waves supposedly propagated through space could not be demonstrated when tested in the experiments of Morley in 1881 and Michelson–Morley in 1887. Studies of blackbody radiation, atomic spectra, and the photoelectric effect had produced observations that could not be explained by classical treatments. These experiments acted as the signposts for modification of the formalisms in the model-making process. The so-called "ultraviolet catastrophe" was a particularly important failing of classical theory because it leads to the discovery of the fundamental quantum constant, Planck's constant. We will examine this case in some detail to gain an appreciation for the thinking that led to the quantum theory.

8.3.1 Thermal Radiation

All objects or bodies emit radiation as a result of their temperature. All bodies absorb and emit radiation from and to their environment. If a body is hotter than its surroundings, it will cool off because the rate of its energy emission will be greater than its rate of absorption. Thermal equilibrium is reached when emission and absorption rates are equal. A human observer can see radiation only in the relatively restricted visible spectrum; so most bodies are visible only because of reflected light. However, if made hot enough, a body will become self-luminous. An example is the thermal radiation generated when a heating coil or an electric stove element is turned on. As the element heats up, the emission of infrared radiation can be detected, felt by the hand but not seen by the eye. Yet, if the element is heated to a higher temperature, it will continue to emit infrared radiation and will also start to glow a dull red, thus now emitting in the visible spectrum as well. Even when an object becomes self-luminous most of the radiation is still in the infrared spectrum. With increasing temperature the element will glow bright red, then white, and finally blue. The change in color with increasing temperature is caused by the shift in the frequency distribution of the emitted radiation. It is the relationship between the frequency distribution of the radiation and the temperature of the body that is the concern of thermal radiation studies.

8.3.2 Blackbody Radiation

In varying degree, the specific spectrum emitted is dependent on the composition of the body. There is a unique class of objects that emit thermal spectra of a uniform or universal nature. These objects are called *blackbodies* because they have surfaces that absorb all incident radiation. At temperatures where they are not self-luminous, they are black. A good approximation to such a black body is an object covered in lamp black. The distribution of the emitted spectrum is universal for all blackbodies thus giving them theoretical importance. The universal character of these distributions allows blackbodies to be treated as an experimental simplification. Thus certain fundamental aspects of the radiation field can be examined with the results of experiments applied generally rather than as a specific experimental case.

At a specific temperature, a blackbody radiator emits a spectral distribution named the *spectral radiancy*, $R_{T(v)}$. This quantity is defined so that $R_{T(v)}dv$ is equal to the energy emitted per unit time in radiation of the frequency defined by the interval $v + dv$ normalized to a unit area of the surface at absolute temperature T.

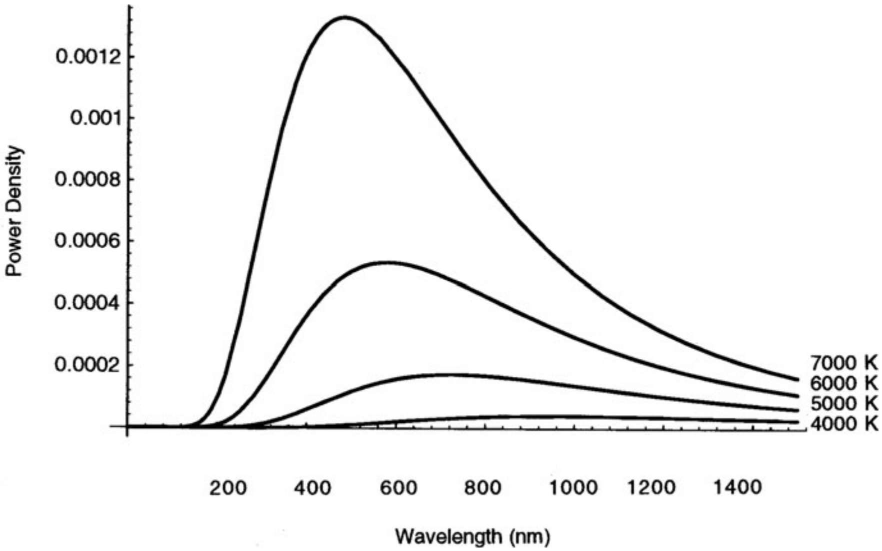


Fig. 8.1 Spectral radiancy of a blackbody radiator at various temperatures

Experimental curves for the relationship of $R_{T(v)}$ on T and v are shown in Fig. 8.1. This family of curves looks very much like the Maxwellian speed distribution function. Inspection of these distributions leads to several observations concerning the power radiated in a fixed frequency interval dv:

1) At very small values of v (with respect to 10^{14} Hz), very little power is radiated. The power radiated at $v = 0$ is 0.
2) The power radiated increases rapidly as v increases from small values.
3) The power rises to a maximum (i.e., it is most intense) at a specific frequency and then falls slowly back toward zero. It is zero as v approaches infinitely large values.
4) The frequency at which the maximum intensity is found increases with increasing temperature. This frequency increases linearly with temperature.
5) The total power radiated (given by the area under a curve at that temperature) increases faster than linear with increasing temperature.

The total energy emitted per unit time from a unit area of a blackbody at a specific temperature is the integral of all $R_{T(v)}$ over all v and is called the *radiancy*, R_T. As discussed above, the relationship between R_T and T is non-linear and is given by *Stefan's law,* formulated in 1879:

$$R_T = \sigma T^4 \tag{8.1}$$

where σ is the Stefan–Boltzmann constant, $\sigma = 5.67 \times 10^{-8}$ W/m^2-K

The frequency at which the intensity maximum is found does increase linearly with temperature and is given by *Wien's displacement law*:

$$v_{\max} \propto T \tag{8.2}$$

or in terms of wavelength

$$\lambda_{\max} T = \text{constant} \tag{8.3}$$

where the constant can be found to have a value of 2.898×10^{-3} m K. This law can be used to determine the temperature of a body (or a star) by measuring its spectral distribution. Wien's law is quite accurate at high frequencies but deviates from experimental values at low frequencies.

A very important blackbody, both historically and experimentally, is a hollow sphere with a small hole or a pinhole in its surface. All radiation incident to the pinhole will enter the cavity. The radiation will be reflected between the walls of the cavity and be absorbed by the walls. An interior thermal equilibrium will be reached in which the walls and the cavity space are radiating and absorbing in steady-state equilibrium. Since the area of the hole is very small with respect to the surface area of the object, no significant radiation escapes from the hole. For all practical purposes therefore, all radiation that strikes the hole is absorbed and consequently *the* hole *has the properties of a blackbody*. In fact, this is the design of most laboratory blackbodies. Consider the converse of this blackbody device. If the walls of the cavity are heated, they will emit thermal radiation that will fill the cavity, again, reaching an equilibrium state. A certain percentage of the cavity radiation will escape from the hole, thus making the hole an emitter of radiation having the properties of a blackbody surface. The hole must emit a blackbody spectrum and, since it is simply sampling the cavity radiation, this means that the cavity itself is filled with a blackbody spectral distribution whose character is dependent on the temperature of the cavity walls. The spectrum emitted by the hole is the energy flux, $R_{T(v)}$, and is proportional to the energy density (the energy contained in a unit volume of the cavity at T in the frequency interval v to $v + dv$), $\rho_{T(v)}$, of the spectral distribution of the cavity radiation:

$$\rho_{T(v)} \propto R_{T(v)} \tag{8.4}$$

Rayleigh in 1899 and then Jeans used a classical approach to calculate $\rho_{T(v)}$.

8.3.3 Classical Theory of Cavity Radiation

In classical terms, radiation is the result of the accelerated motion of a charged particle. Therefore we know that the radiation present in the cavity is a result of the accelerated motion – caused by thermal agitation – of charged particles (electrons) in the walls of the blackbody. Rayleigh and Jeans assumed that the waves inside the three-dimensional cavity exist as standing waves with nodes at the metal surfaces. Using the principle of equipartition of energy, the average kinetic energy of each molecule (or entity) per degree of freedom is ${}^{kT}/_{2}$. Because the standing waves are in thermal equilibrium, and the total energy of an electromagnetic wave is the sum of the electrical and magnetic components, the average total energy of each standing wave according to this formulation is equal to *kT*. This approach yields the important result that the total energy, $\overline{E}$, has the same value for all standing waves in the cavity, independent of their frequencies.

The energy per unit volume in the frequency interval, $v + dv$, in the cavity at T is the product of the average energy times the number of standing waves in the interval divided by the volume of the cavity which gives the Rayleigh–Jeans formula:

$$\rho_{(v.T)}dv = \frac{8\pi v^2 kT}{c^3}dv \tag{8.5}$$

Figure 8.2 shows a comparison of the result predicted by the Rayleigh–Jeans formula and the experimental results from a blackbody radiator. The Rayleigh–Jeans formulation is experimentally valid only in the limit of very low frequencies. However, as the frequency becomes large, the formulation gives the absurd result that the energy densities become infinite. This prediction is itself impossible but furthermore classical treatments cannot explain the experimental picture in which the energy densities remain finite and even fall to zero at high frequencies. This infinite energy source that was dubbed the *ultraviolet catastrophe*.

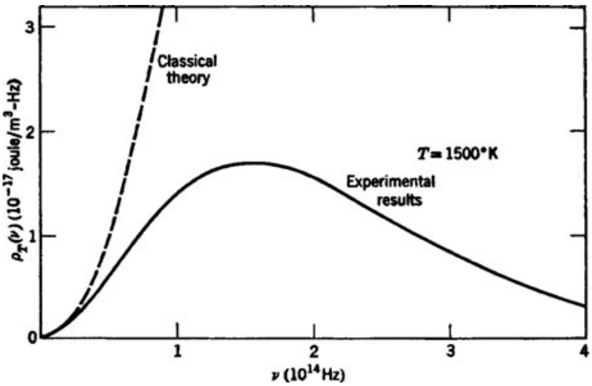


Fig. 8.2 Comparison of the predicted blackbody spectrum as calculated by the Rayleigh–Jeans law and compared to the experimental results. Note that as very low frequencies are approached, the Rayleigh–Jeans law is quite accurate but at higher frequencies it is completely incorrect. This bifurcation in state space is the ultraviolet catastrophe (From *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edition, By Robert Eisberg and Robert Resnick. Copyright by Wiley, 1985. Reprinted by permission of John Wiley & Sons, Inc.)

8.3.4 Planck's Theory of Cavity Radiation

Planck attempted to resolve the ultraviolet catastrophe by considering the cases in which the "law of equipartition" of energy might be violated. The ultraviolet catastrophe occurs because the contribution of the high-frequency oscillators in the blackbody is overstated. Remember that the radiation arises from the accelerated movement of charged particles in the walls of the cavity. Planck sought to reconsider these contributions and his treatment led to a general formula that in one limit becomes the Wien displacement law (the high frequencies). At the other extreme it results in the Rayleigh–Jeans law (the lower frequencies). Planck's treatment actually followed the Rayleigh–Jeans treatment quite closely except that he rejected the assignment of the mean energy for all standing waves as given by the equipartition theory. Instead of the rule, $\overline{E} = kT$, Planck suggested that every energy step be made distinct and each harmonic oscillation be an energy step proportional to the frequency. Thus $\overline{E} = kT$ was replaced by a new expression in which $\overline{E}$, the energy of an oscillator of frequency v, was no longer allowed to be continuously variable but instead was restricted to integral multiples of a new quantity, hv. This quantity is a quantum of energy, and h is a new universal constant now called *Planck's constant*. This modification of classical theory is in fact counterintuitive to common experience and strikes at the heart of one of the fundamental assumptions of classical physics, namely, that a system may take any value in a continuous fashion. Indeed, a charging elephant or a racing bullet clearly moves in a continuous fashion.

The problem of the ultraviolet catastrophe occurs because the equipartition theorem predicts an equal energy assignment to each standing wave generated in the cavity radiator. Standing waves must fit exactly into the cavity. There is a restriction of the wavelengths that can fit into cavity at low frequencies (longer wavelengths). On the other hand, there is absolutely no restriction to the wavelengths that can be fit into the cavity at higher frequencies, for on a continuous scale a shorter wavelength than the preceding can always be made to fit into the cavity. Thus there will be far more waves of high frequency in the cavity. The equipartition law says that all of these waves get an equal amount of energy. Since there are a virtually infinite number of short wavelengths available, the short-wavelength or high-frequency radiation rapidly approaches a total energy content of infinity. That is the Rayleigh–Jeans spectrum.

Planck did not question the treatment of the radiation in the cavity as standing waves and in fact this approach is still generally considered valid. Planck recognized that at the restricted low frequency end of the spectra where $\overline{E} \rightarrow kT$ as $v \rightarrow 0$, the Raleigh–Jeans law was accurate. What is needed is a high-frequency cut-off such that $\overline{E} \rightarrow 0$ as $v \rightarrow \infty$. Planck reasoned that if the energies were dependent on the frequency, the experimental results could be modeled in theory. However, this approach could come only at the expense of the equipartition law, which clearly states that energy is independent of frequency. Planck leapt into the unknown. By embracing a discrete rather than continuous phenomenology, the integration of the Boltzmann distribution on which the equipartition theorem was based was changed to summation of the distribution. He postulated the following rule to replace the equipartition theorem.

The energy of a harmonic oscillation will be an integer multiple of an energy step which is proportional to the frequency:

$$\overline{E} = nh\nu \quad \text{where} \quad n = 1, 2, 3\text{K} \tag{8.6}$$

Thus the energy of the intervals ΔE between frequencies becomes restricted, and a frequency dependence for $\overline{E}$ is the result. The simplest relationship that provides this result is the proportionality of ΔE and ν, which can be written as an equation:

$$\Delta E = h\nu \tag{8.7}$$

The proportionality constant in Eq. (8.7), h, is Planck's constant. Planck calculated h from a best fit of theory with data and arrived at a value quite close to the presently accepted number. He stated at the time that his ad hoc approach would be verified as correct only if this constant were found to apply to other problems and situations in physics. This verification was soon to appear.

Using the relationship described in Eq. (8.7) and summation rather than integration in the Boltzmann relationship yields Planck's blackbody spectrum:

$$\rho_T\,(\nu)d\nu = \frac{8\pi\nu^2}{c^3}\frac{h\nu}{e^{h\nu/kT} - 1}d\nu \tag{8.8}$$

The spectrum predicted by this model shows impressive agreement with experimental blackbody spectra. Furthermore, in the classical limit, Planck's formula reduces to the Rayleigh–Jeans law, and at high frequencies it leads to Wien's displacement law. These quite spectacular results are the consequence of a single and "simple" adjustment of the model: instead of giving entities access to a continuous set of energy levels access is restricted to a discrete series of values.

8.3.5 Quantum Model Making – Epistemological Reflections on the Model

The quantization expressed in Planck's treatment of the blackbody radiation problem can be generalized: the relation given in Eq. (8.6) applies to any physical entity with a single degree of freedom that executes simple harmonic oscillations (i.e., the coordinate, or instantaneous condition, of the degree of freedom is a sinusoidal function of time). Thus the length of a spring, the angle of a pendulum bob, and the amplitude of a wave are all subject to this postulate. These are objects of common experience, and for the most careful and fully schooled observer has difficulty imagining that the motion of grandfather's clock pendulum is anything but continuous in the sense of classical physics. It seems almost axiomatic that grandfather's clock pendulums are not subject to discrete jumps in energy. Discontinuous movements

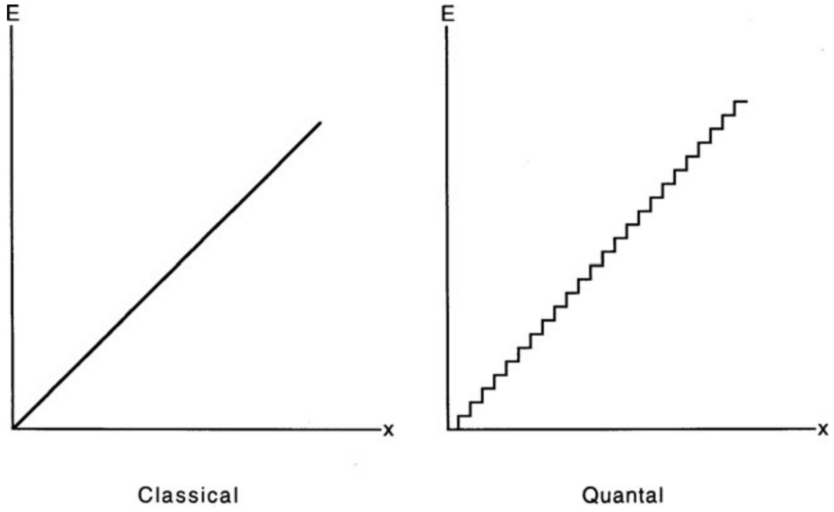


Fig. 8.3 The possible energy states that can be taken by an entity is continuous in the classical treatment but limited to discrete energy levels in the quantum treatment

of the bob are never observed. How can such a conflict be understood? Consider Fig. 8.3. If the observed events and their associated energies are considered classically, the first graph describes the continuum of energy levels to which this harmonic oscillator has access. In contrast, the discrete jumps of the quantized system are indicated in the companion graph. The key to resolving this conflict is to consider the system quantitatively and not qualitatively.

Question: Is the energy loss due to friction in a pendulum a continuous or discontinuous (discrete) process?

We will consider a pendulum consisting of a 0.02 kg mass suspended from a string 0.2 m long. The amplitude of the oscillation will be the maximum angle of deviation from the vertical, 0.1 rad.

We will calculate the energy needed for each discrete quantum jump for the pendulum using: $\Delta E = hv$ and compare that energy to the total energy of the pendulum given by its maximum potential energy, $E = mgh_{(\text{height})}$. The oscillation frequency, v, is given by:

$$\begin{aligned} v &= \frac{1}{2\pi}\sqrt{\frac{g}{l}} \\ &= \frac{1}{2\pi}\sqrt{\frac{9.8\,\text{m s}^{-2}}{0.2\,\text{m}}} = 1.114\,\text{s}^{-1} \end{aligned} \tag{8.9}$$

The energy of the pendulum is

$$\begin{aligned} E &= mg\,[l\,(1-\cos\theta)] \\ &= (0.02\,\text{kg})(9.8\,\text{m s}^{-2})[\,0.2\text{ m}(1-\cos 0.1)] \\ &= 1.96\times 10^{-4}\text{J} \end{aligned} \tag{8.10}$$

The jumps in quantized energy are

$$\begin{aligned}\Delta E &= h\nu \\ &= (6.63 \times 10^{-34}\ \text{j s})\ (1.114\ \text{s}^{-1}) \\ &= 7.39 \times 10^{-34}\text{J}.\end{aligned} \tag{8.11}$$

Comparison of the result in Eq. (8.10) with that of Eq. (8.11)

$$\frac{\Delta E}{E} = \frac{7.39 \times 10^{-34}\text{J}}{1.96 \times 10^{-4}\text{J}} = 3.77 \times 10^{-30} \tag{8.12}$$

indicates that in order to demonstrate that the energy loss is discrete would require a measuring instrument with a resolution of approximately 4 parts in 10^{30}!

Since there are no measuring instruments capable of the measurements required to answer the question, it is clear that ordinary pendulums cannot be used to determine the validity of the Planck quantum postulate. It is the smallness of the Planck constant that makes the discrete graininess of the Universe appear no different than a continuum of observables in all macroscopic systems. For all practical purposes, h can be zero for systems in the classical limit. Thus for planets, bowling balls, and charging tigers, a continuum is a perfectly reasonable abstraction for the observed state space. This is the explanation for the conflict expressed above. We can predict that the discrete effects of the Planck postulate will become observable when ν is large or when E becomes very small (as with objects of very small mass). As we will see in many of the topics that follow, in matters of interest to the biologist and biophysical chemist, there are many cases in which the classical limit can be properly invoked, but there will also be numerous cases in which an understanding of the quantum rules will be necessary.

8.4 The Concept of Heat Capacity Was Modified by Quantum Mechanical Considerations

Planck's constant and his quantum postulate found application in a very similar way to the problem of an observed variation in heat capacity. In 1819, Dulong and Petit performed experiments at room temperature on a variety of solids and found that the heat capacity, c_v, was approximately 6 cal/mol K regardless of the material tested and that the heat capacities were constant. Thus the heat required to raise the temperature of a solid in a given amount was apparently independent of the elements of composition and of the temperature.

This experimental result was supported by a classical statistical treatment using the equipartition theorem. However, further experiments, at lower temperatures, showed that all molar heat capacities vary and tend toward zero as the temperature is decreased. Near absolute zero, the heat capacity varies as T^3. Einstein had recognized the importance of Planck's quantum hypothesis and, in 1905 first applied the black-body radiation result to the photoelectric effect; then in 1906 extended this work to the specific-heat problem. Einstein recognized the kT factor as a remnant from equipartition theory that had to be replaced by a new factor that took into account the quantization of a harmonic oscillator. In this treatment, a temperature dependence is found, and the curves agree qualitatively with experiment except that a different oscillator frequency has to be chosen for each element to get agreement with experiment. Furthermore, Einstein's treatment does not account for the T^3 dependence at low temperatures. The application of the quantum postulate therefore is qualitatively quite successful but not adequate to account for the observed behavior of solid material in the phase space.

Debye used a different theoretical approach to the question and successfully calculated the exact experimental results. In departing from Einstein's treatment, Debye considered the particles in the solid to be interacting and hence allowed a distribution of frequencies due to the collective oscillation of the molecules in the crystal. The crystal will oscillate in a distribution of frequencies that are dependent on the rigidity of the crystal, up to a maximum of v_{max}. This deviation is expressed as *Debye's* T^3 *law* and is used to calculate entropies at low temperatures in thermodynamic problems.

8.5 The Photoelectric Effect and the Photon-Particle Properties of Radiation Could Be Understood Using Planck's Quanta

Following Planck's quantum treatment of the black-body radiation problem, the quantum postulate was used successfully in reworking several of the disconsonant problems of classical physics. We have seen two cases so far in which classical treatments could not even qualitatively explain the experimental results. Several more areas of classical confusion were to yield quickly to the quantum postulate in the first years of the twentieth century. Like the very waves that were under reevaluation, the effects of the quantum postulate rippled and broke on the formerly calm shores of many fronts of physics and chemistry. The quantum postulate implied a particulate graininess in the Universe and these effects could not be accounted for in the wave mechanics of classical treatments. The restriction to discrete energy states, although it explained a variety of physical phenomena, also implied particle-like qualities in waves and wave-like qualities for particles. Atomic structure and chemical activity became bound by certain restrictions and rules. The atom was no longer an indivisible and stable fundamental particle. As we look back through the temporal lens we recognize that with greater and surer understanding came a new fundamental uncertainty about what could actually be known!

Albert Einstein is a central player in the application of Planck's quantum postulate to a number of classical conundrums. Before he applied the quantum treatment to the heat capacity problem, he had considered the results of a confusing series of observations made by Heinrich Hertz in 1887. Hertz was experimenting with spark gaps. He noted that a spark jumped more readily between two electrified gap electrodes when that gap was illuminated by the light emanating from another spark gap than when the sparking of a gap electrode was kept in the dark. Study of this phenomenon over the next two decades made it clear that electrons were emitted from surfaces of solids with a dependence on the frequency of the incident light. The kinetic energy of the photoelectrons can be determined by varying an electric potential, Φ, that retards the ejected photoelectron (Fig. 8.4). Photoelectrons with KE $< \frac{1}{2}mv^2 = e\Phi$ will not reach the detector target, thus allowing the maximum kinetic energy of the photoelectrons to be determined. In 1902 P.E. Lenard showed that the maximum kinetic energy depended on the frequency of the incident light regardless of the intensity of the light. Below a certain frequency v_0, no electrons were ejected regardless of the intensity of that light (Fig. 8.5). Above v_0 the photoelectrons were ejected as soon as the incident light hit the surface even if the intensity was exceedingly feeble.

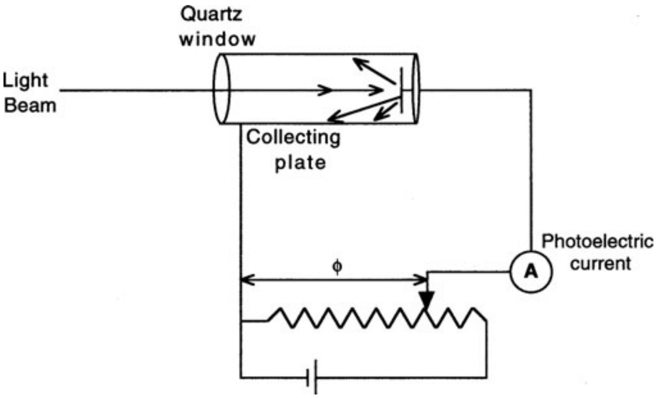


Fig. 8.4 Diagram of the apparatus used to study the photoelectric effect

These observations could not be explained by the classical wave theory of light, and Einstein proposed a solution based on a quantization of the radiation field. This quantization contends that the radiation field accepts or supplies energy in discrete amounts. In the case of the photoelectric effect, the energy is transferred completely to an electron overcoming the binding energy of the material. The amount of energy required to free the electron from its atomic binding is called the *work function*, Φ, and any excess energy from the radiation field will appear as kinetic energy of the photoelectron:

$$\frac{mv_o^2}{2} = hv - \Phi \tag{8.13}$$

A quantized amount of energy less than the work function will not generate a photoelectron. This is precisely the behavior described in Fig. 8.5. Einstein thus applied the idea of quantized energy delivery as the explanation for the photoelectric effect.

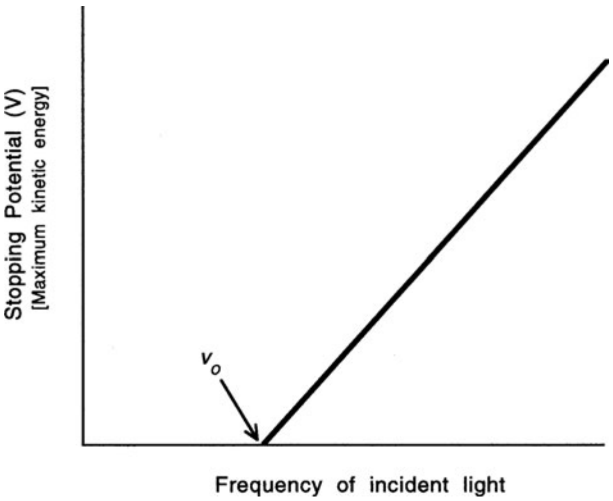


Fig. 8.5 Regardless of the intensity, photoelectrons are ejected only when light above the frequency v_0 illuminates the surface

8.6 Electromagnetic Radiation Has a Dual Nature

Though the photoelectric as well as observations by A.H. Compton on the scattering interactions of a beam of electrons and gamma ray light make it clear that electromagnetic radiation is composed of particles, the observed electromagnetic phenomena of diffraction and interference also require a wave theory of radiation. Thus electromagnetic radiation is both particle-like and wavelike and behaves like one or the other depending on the circumstances. In fact, Compton's experiments are a testament to this dual nature (see Appendix H). The measurements of wavelength were made by a crystal spectrometer, and they were interpreted as a diffraction phenomenon even though the scattering data can be interpreted only by treating the gamma light as particulate in nature. If light so well described in many cases as a wave has particle-like properties, what is the likelihood that particles such as electrons and protons could have wavelike properties? It turns out that wave-particle duality is a thoroughly general phenomenon in nature and that all entities have wave and particle qualities. In other words, in natural state space, all known entities have wave-particle duality, and their appearance as a particle or wave is a result of the observables chosen to describe the system. In many cases this leads to a strong interaction between observer and the natural state space, which means that the observer is an interactive participant in the state space. This condition of drawing the observer into the state space as a part of the natural system substantially changes what we can know about the system. The maximum knowledge that can be acquired from a natural system is described by the *indeterminacy principle* put forth by Werner Heisenberg in 1927. We will explore this principle in some detail shortly, but for now we will recognize that it codifies the effect of the observer–state space interaction. It turns out that classical formulations deterministically predict a greater knowledge about a natural state space than is possible because the classical approach underestimates the interaction between the observer and the system. There is always a minimal interaction and the magnitude of this interaction is related to Planck's constant. It is vital to realize that this interaction is *not* an experimental error that can be eliminated by more careful measurement or a more clever experimental apparatus. The interaction between observer and systematic state space is itself part of the system and cannot be "corrected" or nullified.

8.7 de Broglie's Postulate Defines the Wavelike Properties of Particles

de Broglie noticed a mathematical connection between the formulation by Fermat that explained the path of waves of light in optical systems, *the principle of least time*, and a similar formulation by Hamilton explaining the path of particles in mechanical systems, the *principle of least action* (see Appendix I). This connection led him to propose that any moving body has an associated wave and that there is a relationship between the momentum of the particle and the associated wavelength. This relationship is the deBroglie relation:

$$p = \frac{h}{\lambda} \tag{8.14}$$

This relationship, proposed in 1924, codifies the *duality* of the nature of things. All things have both wavelike and particle-like qualities. The issue becomes which of these opposite properties is most useful as an observable in describing the system in which the thing exists. The photoelectric effect provided evidence that light, which is very wavelike also has a particulate nature. Does similar evidence exist that particles have wavelike behaviors? de Broglie's conjecture could be tested: Could a stream of particles be diffracted? Actually this experiment had already been done in 1921. An experiment by Davisson had demonstrated an anomalous scattering peak, but its meaning had not been appreciated. In 1925 Davisson and Germer performed another experiment in which electrons were diffracted quantitatively in agreement with the wavelength predicted by the de Broglie relation. Finally in 1927, G.P. Thomson demonstrated electron diffraction in thin films and won the Noble Prize for this work (in 1937, along with Davisson). An oft-repeated historical note by Max Jammer is that G.P. Thomson was the son of J.J. Thomson. J.J. Thomson won the Noble prize in 1906 for his discovery of the electron which he described as a particle with a defined mass-to-charge ratio. Jammer wrote, "One may feel inclined to say that Thomson, the father, was awarded the Noble prize for having shown that the electron is a particle and Thomson, the son, for having shown that the electron is a wave." The practical value of matter waves is used daily in the electron microscope and in neutron diffraction studies. All material objects regardless of charge or

Table 8.2 Wavelengths associated with various particles

Particle	Mass (kg)	Velocity	Momentum (m s–^{1})	Wavelength (m)
Electron	9.11×10^{-31}	3×10^5	2.73×10^{-25}	2.43×10^{-9}
Electron	9.11×10^{-31}	300.00	2.73×10^{-28}	2.43×10^{-6}
Proton	1.67×10^{-27}	3×10^3	5.01×10^{-24}	1.32×10^{-10}
Deuteron	3.34×10^{-27}	3×10^3	1.00×10^{-23}	6.626×10^{-11}
Carbon atom	1.99×10^{-26}	300.00	5.97×10^{-24}	1.11×10^{-10}
Water molecule	2.99×10^{-26}	300.00	8.97×10^{-24}	7.39×10^{-11}
Myosin	7.31×10^{-22}	300.00	2.19×10^{-19}	3.02×10^{-15}
Baseball	0.136	44.78	6.09×10^0	1.09×10^{-34}
Speeding bullet	0.030	300.00	9.00×10^0	7.36×10^{-35}
Charging elephant	7000	7.5	5.25×10^4	1.26×10^{-38}

size show wavelike characteristics in their motion (Table 8.2). For massive objects, however, the wavelength is so short that the wavelike properties are easily missed and their particulate properties dominate. It is not until particles with relatively long wavelengths such as electrons and protons are considered that the wavelike properties are too prominent to ignore. These properties are a consequence of the small but non-zero value of *h*.

8.8 The Electron Microscope Employs Particles as Waves to Form Images

The electron microscope takes advantage of both the charge and the wave nature of a high-velocity electron. Like x-ray diffraction, the very small wavelengths allow extremely high resolution, but as in the light microscope, the charge on the electron allows the electron beam to be collimated and focused by electromagnetic lenses (Fig. 8.6). The wavelength limit to resolution of the optical microscope originally led to an attempt to improve resolution by using shorter wavelengths. For example, in the ultraviolet microscope, $\lambda = 250$ nm and so the resolution $\approx$ 110–120 nm. Much shorter wavelengths are possible using electrons. The wavelength obtained at a given accelerating voltage, *V*, can be calculated from

$$\lambda = \frac{12.3}{v^{1/2}} \tag{8.15}$$

The constant in the numerator is derived from the physical constants for Planck's constant and the mass and charge of an electron. For an accelerating voltage of 50,000 V [a common potential used in commercial electron microscopes] the wavelength will be 0.054 Å. Because the microscope is operated under a vacuum, the

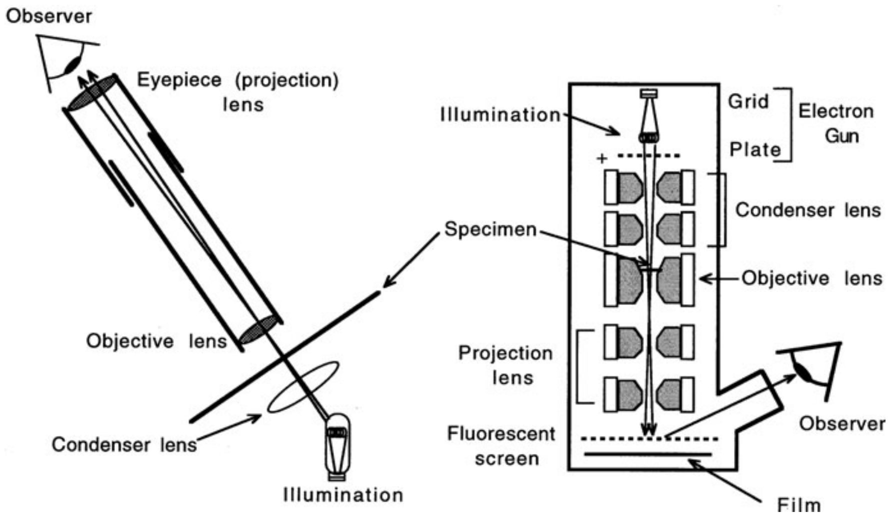


Fig. 8.6 Comparison of the lens arrangements of the light microscope (*left*) and the electron microscope (*right*)

refractive index of the medium does not enter into the resolution calculation. The practical requirements to limit spherical aberration make the value of α small so cos $\alpha = \alpha$ thus, resolution, d:

$$d = \frac{0.611}{\alpha} \tag{8.16}$$

The angle of illumination is usually near 4×10^{-3} radians thus making the theoretical resolution of a 50 kV microscope 0.01 Å. The actual resolution achieved is between 5 and 10 Å. This over 500-fold degradation in theoretical resolution is almost entirely due to spherical aberration of the magnetic lenses. The spherical aberration makes the image of a focused point spread into a disc. The diameter of such a disc limits the resolving power dramatically. Unlike optical lenses, the spherical aberration of magnetic lenses is not easily corrected, so at the present time only a reduction in *numerical aperture* can limit the loss in resolution due to aberration:

$$d = Cf\alpha^3 \tag{8.17}$$

where C is a constant and f is the focal length of the objective lens. Obviously Eqs. (8.16) and (8.17) are in conflict in terms of resolution, and the practical result is a compromise between the two.

In the light microscope an image is seen because of absorption of the light by the sample. But in the electron microscope the image is formed because of scattering of the electrons by interactions between the beam electrons and the specimen atoms. Scattering of the electron beam is both elastic and inelastic. Elastic scattering occurs when the path of the electron is altered but its velocity is not affected. The wavelength of these electrons is not affected and they can be focused, thus forming an image. Inelastic scattering is said to occur when both the path and the velocity of the beam electron are changed. Inelastic scattering does not contribute to image formation because the alterations in velocity change the wavelengths of the electrons, which are then focused on different image planes and are only seen as background fog. The degree of deflection of a beam electron depends on the nuclear charge of the specimen atom, the larger the positive charge, the greater the deflection. This is the reason why heavy atoms such as osmium, uranium, and lead are effective electron stains in biological samples. The imaging of molecules, such as polypeptides and nucleic acids, and fresh biological samples after freeze-fracture is accomplished by a process called *shadowcasting.* Shadowing occurs when a metal such as carbon, gold, or platinum is evaporated in a high vacuum thus spraying a gas of metal atoms onto a sample. The sample thus has enhanced contrast.

8.9 The Uncertainty Principle Is an Essential Conclusion of the Quantum Viewpoint

The *uncertainty* or *indeterminacy* principle is a consequence of the non-deterministic and probabilistic viewpoint of quantum physics. The apparent determinism seen in macroscopic systems such as cannonballs and planets, which allows the simultaneous knowledge of momentum and position to any arbitrary degree of accuracy, is an incorrect though usually permissible approximation of the real world. This abstraction is not useful in the microscopic world. A problem occurs because there is need to measure two *complementary* observables simultaneously. Heisenberg and Bohr considered this problem and recognized that there is a limit to the ability to know certain types of information simultaneously, namely momentum and position, and time and energy. We can write the limit as

$$\begin{aligned} \Delta \mathbf{p}_x \Delta x &\geq \frac{h}{4\pi} \\ \Delta t \Delta E &\geq \frac{h}{4\pi} \end{aligned} \tag{8.18}$$

The *Heisenberg indeterminacy principle* defines the degree of accuracy with which we can know both properties at the same time. We can know precisely the position of the particle but consequently we cannot be confident at all of knowing its momentum. The restriction is on the product of the properties in a simultaneous measurement. Consider the problem in the following thought experiment. Bohr proposed using a γ-ray microscope to view an electron. We wish to pinpoint the position of an electron by observing it under a microscope. In order to see the electron, we must illuminate it since we actually "see" the electron by detecting a light photon scattered by it. The act of illuminating the electron disturbs it, and it recoils according to the Compton effect, in a somewhat unpredictable manner. The uncertainty principle is in effect. If we don't illuminate it, we can't see it at all! This is the coupling of the observer to the system that was referred to earlier. The act of measuring itself contains a degree of indeterminacy that is part of the nature of the observation. How is the momentum of the particle related to the determination of the particle's position? Assume the intensity of the illuminating light is low enough that only one photon enters the objective lens. The momentum of the scattered photon is $p = {}^{h}/_{\lambda}$ but it may have been scattered anywhere within the angular range of 2θ subtended by the objective lens. The x-component of the momentum varies between $\pm p \sin\theta$. Thus the uncertainty is

$$\Delta \mathbf{p}_x = 2\sin\theta = \frac{2h}{\lambda}\sin\theta \tag{8.19}$$

Because of conservation of momentum the electron will receive a recoil momentum in the x direction equal in magnitude to the x momentum change of the photon; hence the x momentum of the electron is equal to Eq. (8.19). Perhaps we can do better with the location of the electron along x. The problem is that a microscope's image of a point object is a diffraction pattern therefore the image of the electron is going to be blurred. The resolving power of the microscope will decide the accuracy of positional determination. If we use the central diffraction maximum as the uncertainty in x along with the formula for the resolving power of a microscope, we can write

$$\Delta x = \frac{1}{\sin\theta} \tag{8.20}$$

Thus the photon originated somewhere in this x axis range. We can show that the product of $\Delta \mathbf{p}_x \Delta x$ is in the range of Eq. (8.18):

$$\Delta \mathbf{p}_x \Delta x = \frac{2h}{\lambda}\sin\theta\left(\frac{\lambda}{\sin\theta}\right) = 2h \tag{8.21}$$

Note that we could reduce $\Delta \mathbf{p}_x$ by lengthening the wavelength or using an objective lens with a smaller numerical aperture. Alternatively we could reduce Δx by shortening the wavelength and using an objective with a larger numerical aperture. Obviously we cannot simultaneously improve both uncertainties.

8.10 An Historical Approach to Understanding Atomic Structure and the Atom

Coincident with the experiments that showed the Universe to be lumpy with a graininess quantified by Planck's constant, a link between atomic structure and atomic spectra was being developed. Early models were inadequate until Niels Bohr added the quantum graininess to the models. Again for our purposes of seeing model-making in action, we will take a historical path to an understanding of atomic structure.

It is difficult to imagine in the early years of the twenty-first century, surrounded by iPhones, iPods, MRIs, and lasers, that the focus of physical science until the early decades of the twentieth century was mechanics and not electricity, optics, or magnetism. There could not be great progress in understanding the atom without an appreciation of its electrical nature and of the relationship between electricity, magnetism, and light. The early unification theory of electricity, magnetism, and light by Maxwell in 1865 allowed light to be explained as an electromagnetic wave. However, this theory could not explain the observed spectral lines of atoms or the fact that atoms do not lose all their energy when they radiate light. Hertz had confirmed Maxwell's theory in 1887 by generating radio waves from sparks. The seminal work of Michael Faraday on electrolysis established that the forces holding atoms together were electrical. Faraday's work established that ionic charges are integrals of charge and provided a measure of the ratio of the masses of atoms to the electrical charges of the ions. He suggested the electrical nature of matter and the existence of subatomic particles as well as the existence of a fundamental unit of charge. Faraday wrote

> The atoms of matter are in some way endowed or associated with electrical powers, to which they owe their most striking qualities, and amongst them their chemical affinities

The existence of the electron was established in 1897 by J.J. Thomson following up on experiments with cathode rays by H. Geissler and Plücker. These workers had improved the vacuum tube and forced current through two sealed electrodes in the tube. In the first experiments a green glow attributed to rays emanating from the cathode were noted (1858). Later a student of Plücker 's, J.W. Hittorf demonstrated the shadow of an object placed between the cathode and a screen, thus establishing that the cathode rays indeed came from the cathode (1869). Then in 1879, W. Crookes showed that cathode rays could be bent by a magnetic field in a direction that suggested that they were negatively charged. He showed that the associated luminescence was independent of the gas in the tube or the metal from which the electrode was fashioned and concluded that the cathode rays were a property of the electric current. Thomson then established that the electron was a negatively charged subatomic particle and determined the charge-to-mass ratio. The existence of positively charged particles thousands of times heavier than electrons (ions of the gas in the tubes) was established by W. Wien and refined by Thomson who was able to establish the existence of isotopes of elements. The technique was further developed by F.W. Aston, who invented the mass spectrograph.

Since it was known that atoms are electrically neutral, the existence of the negatively charged electron posed a significant problem (Fig. 8.7). It meant that there had to be a positive charge to be neutralized. Thomson suggested the "plum pudding" model, in which the electrons were arranged in a regular pattern in a uniformly charged sphere of positive charge. This was an inherently stable model in which any extracted electron would attempt to return to its naturally assigned place. Since classical electrodynamics required the vibration of a charge such as an electron to

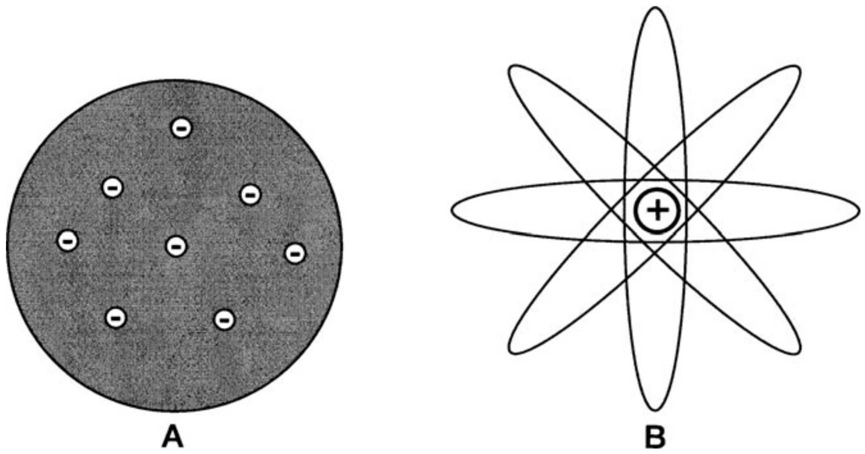


Fig. 8.7 Pre-quantum atomic models: (**a**) Thomson's model; (**b**) Rutherford's model

generate an electromagnetic wave, the Thomson model could qualitatively explain the generation of light by excited atoms. However, such a model cannot explain the observed spectra of atoms. Rutherford disproved the Thomson model with a series of experiments in which he fired alpha particles at ultrathin targets of gold foil. He found that the great majority of the time the alpha particles would penetrate the foil without any deflection. Only occasionally would there be certain particles that would be sharply deflected onto a screen placed to the side of the apparatus. Rutherford described the results,

> It was almost incredible, as if you fired a 15 inch shell at a piece of tissue paper and it came back to hit you. On consideration, I realized that this scattering backwards must be the result of a single collision, and when I made calculations I saw that it was impossible to get anything that order of magnitude unless you took a system in which the greater part of the mass of the atom was concentrated in a minute nucleus. It was then I had the idea of an atom with a minute massive centre carrying a (+) charge.

Rutherford's model was reminiscent of the planetary or "Saturnian" model proposed in 1904 by H. Nagaoka. The difficulty with the Saturnian model is that an electron in constant circular motion is constantly accelerated and should therefore continuously radiate light. This continuous radiation means that the electron will gradually lose energy and must spiral relentlessly into the nucleus. This model is inherently unstable. In classical mechanics a model of the atom as a small positive sphere surrounded by negatively charged particles is impossible; yet experimentally this is precisely what is found. One of the great successes of quantum theory is that it explains the experimental data but at the expense of our relatively comfortable classical world view. Thus we see that our classical common-sense perspective is in reality a linkage based on classical ideas coupled to the observables in the natural (quantum) state space.

8.10.1 Atomic Spectra

From 1814 to 1824 J. von Fraunhöfer observed the Sun through a telescope that he had fitted with a fine optical prism. He observed that the Sun's spectrum was not continuous but instead had hundreds of dark lines, the most prominent of these he labeled A through G: the Fraunhöfer lines. Fraunhöfer did not know the significance of these lines, and 35 years passed before an explanation for them was forthcoming. Kirchhoff in 1844 heated various elements to incandescence and observed the emitted spectra through a spectroscope (Fig. 8.8). Unlike the continuous spectra of a blackbody, the spectra of an element of free atoms, was a series of distinct lines of color (the lines being formed by the image of the spectroscopic slit). Each element had its own characteristic set of spectral lines that could be seen in a spectroscope regardless of chemical combination with other elements. In 1848 Foucault showed that a sodium flame emitting a D line would also absorb this same line if a sodium arc were placed behind it. It was Kirchhoff who related absorption and emission with the law: "The relation between the power of emission and the power of absorption for rays of the same wavelength is constant for all bodies at the same temperature." Thus he summarized the phenomena of spectroscopy: *a body that absorbs light at one wavelength also emits light at the same wavelength*. In 1859 R.W. Bunsen and Kirchhoff discovered cesium and rubidium by observing their spectral lines. In a similar fashion the element helium was discovered in the Sun by spectroscopy long before it was found on the Earth.

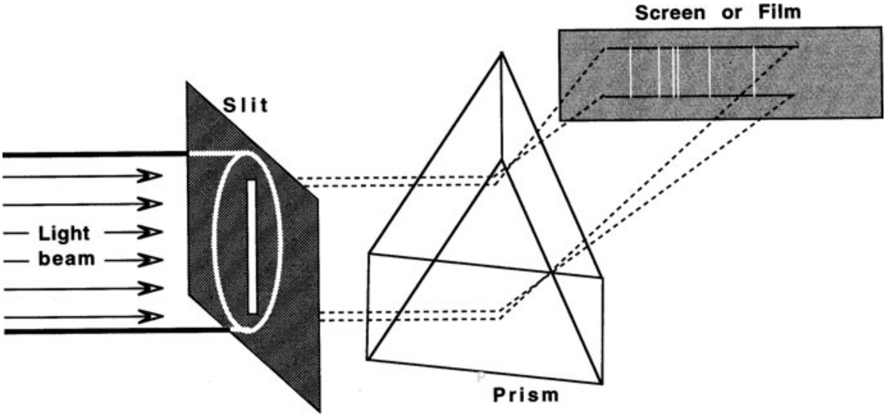


Fig. 8.8 Schematic of the atomic spectroscope

The spectrum of most atoms is extremely complicated with often hundreds of discrete lines. The spectrum of hydrogen is relatively quite simple. Hydrogen is the simplest atom with a single electron and a single proton. The simplicity of its spectrum seems logical. The spectrum of hydrogen is very regular with the spacing of lines, in terms of decreasing wavelength, converging to a series limit at 3645.4 Å (Fig. 8.9). The regular nature of the spectrum led Balmer and others to propose formulas that predicted a number of the lines. Balmer's equation

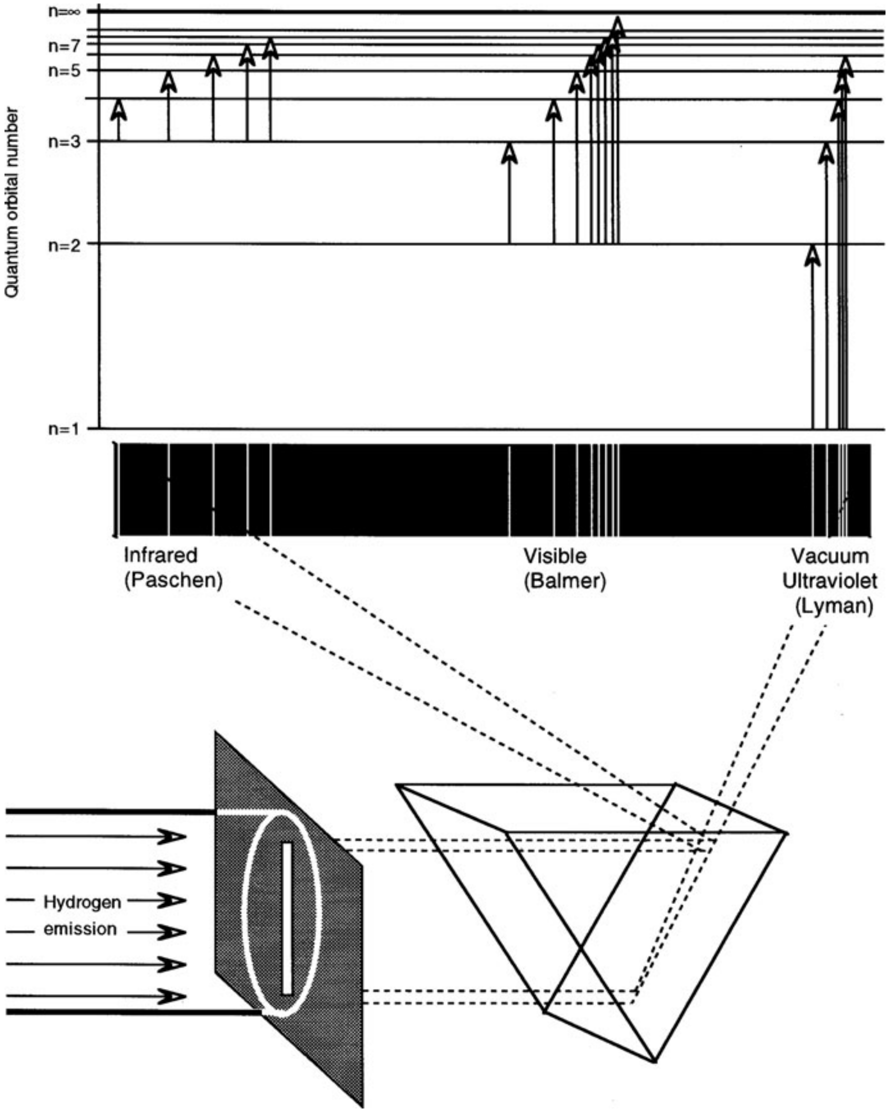


Fig. 8.9 The spectra and transitions of the series for the hydrogen atom. The *arrows* pointing to transitions in this figure indicate the magnitude between an excited energy state and the ground state. When an electron falls down this gradient a photon of specific energy is emitted. Note that the larger the magnitude of the energy difference the shorter the wavelength (and higher energy) of the emitted photon

$$\lambda = 3646\frac{n^2}{n^2-4} \text{ for } n = 3, 4, 5\mathrm{K} \tag{8.22}$$

was able to predict the first nine lines of the spectrum (near ultraviolet and visible). Other series were derived including the Lyman (ultraviolet), Paschen (infrared), Brackett (infrared), and Pfund (infrared) series. The accuracy and detail of spectroscopic measurements set a very stringent limit on the atomic models that could be proposed. The first success in proposing a model that could meet this quantitative goal post was accomplished by Niels Bohr in 1913.

8.10.2 Bohr's Model

Bohr created a model for atomic structure that freely mixed classical and non-classical aspects in an amalgam reminiscent of Planck's treatment of the black-body problem. Bohr recognized that the nuclear model of the atom proposed by Rutherford could not exist in classical theory. The solution to the stability problem was presented as a series of postulates which fixed the size of the atom. This required an equation with a fundamental constant that has length. Charges and masses do not have length so Bohr needed another constant. He noted that Planck's quantum constant yielded a measure of length (action) when combined with charge and mass. The numerical value of this combination was close to the size of an atom, and so Bohr used Planck's constant in making his model. Bohr's postulates are

1) Obeying classical mechanics an electron in an atom moves in a circular orbit around the nucleus under the coulomb attraction between the two.
2) The available orbitals are limited to those in which an electron's orbital momentum L is an integral multiple of ${}^{h}/_{2\pi}$ (i.e., $L = n\mathrm{h}$). (The term ${}^{h}/_{2\pi}$ is so common that a new constant, $\hbar$, is frequently used instead).
3) Despite the constant acceleration of the electron, an electron in a permissible orbit will not radiate electromagnetic energy. E_{total} remains constant.
4) Electromagnetic radiation is emitted only when an electron moving in an orbit with total energy E_i moves discontinuously to another orbit of total energy E_f. The frequency of the emitted radiation v is equal to $\frac{E_\mathrm{i}-E_\mathrm{f}}{h}$.

The first of these postulates assumes the nuclear atom. The electron will be mechanically stable in orbit when the coulomb force acting between the electron and nuclear charge is offset by the centrifugal force (*ma*):

$$\frac{1}{4\pi\varepsilon_\mathrm{o}}\frac{Ze^2}{r^2} = m\frac{v^2}{r} \tag{8.23}$$

Z is the nuclear charge, v is the velocity of the electron in orbit, and r is the radius of the orbit. The orbital angular velocity of the electron is quantized, $mvr = L = nh$. This restricts the possible circular orbits which with some algebra gives

$$r = 4\pi\varepsilon_0 \frac{n^2\mathrm{h}^2}{mZe^2} \text{ where } n = 1, 2, 3\mathrm{K} \tag{8.24}$$

The total energy of an electron moving in one of the allowed orbits can be found by calculating the potential and kinetic energy of the electron:

$$U = -\frac{Ze^2}{4\pi\varepsilon_0 r} \text{ and } KE = \frac{mv^2}{2} = \frac{Ze^2}{4\pi\varepsilon_0 2r} \tag{8.25}$$

thus

$$E = K + V = -\frac{mZ^2e^4}{(4\pi\varepsilon_0)^2\, 2\, h^2}\frac{1}{n^2} \text{ where } n = 1, 2, 3, \mathrm{K} \tag{8.26}$$

The energy change as an electron makes a transition from one orbit to another is

$$\Delta E = \left(\frac{mZ^2e^4}{(4\pi\varepsilon_0)^2\, 2\, h^2}\right)\left(\frac{1}{n_1^2} - \frac{1}{n_2^2}\right) \tag{8.27}$$

This expression can be used to predict the wavelength of the spectral line by using $hv = \Delta E$and $v = \frac{c}{\lambda}$. The Bohr model very accurately calculated the wavelengths of the Balmer series. Furthermore, it predicted the spectral lines for the transitions of the Lyman and other series which had not yet been discovered. The link between atomic structure and spectra is driven home by this analysis of the Bohr model; the quantization of the orbitals is fundamental to understanding the structure and the behavior. The Bohr model is an unusual mix of classical and quantum concepts. It is still often used as an abstraction because it is somehow compellingly simple, but the mixture is fundamentally flawed. Before we move toward a more consistent picture of the quantum structure of the atom let us merge the story lines of atomic structure and wave–particle duality.

In 1924 de Broglie provided a physical interpretation of the quantization rule of the Bohr atom. If the electron is a particle with quantized angular momentum it is also a quantum particle with an associated matter wave. We can relate these properties

$$mvr = pr = n\mathrm{h} \quad n = 1, 2, 3, \ldots \tag{8.28}$$

where p is the linear momentum of the electron in the orbit of radius r. We substitute the de Broglie relation, $p = {}^{h}/_{\lambda}$, into the Bohr equation:

$$\frac{hr}{\lambda} = \frac{nh}{2\pi} \tag{8.29}$$

Then

$$2\pi r = n\lambda \quad n = 1, 2, 3, \ldots \tag{8.30}$$

The interpretation of Eq. (8.30) is that *the allowed orbits will be those in which an integral number of de Broglie wavelengths will fit perfectly*. Imagine the electron and its associated wave wrapped repeatedly around the orbit. The integral fit means that the waves must fit exactly one on top of the other. In other words they are precisely in phase. There is no picture of progressive wave motion but rather an image of standing waves. The number of standing waves depends on the length of a specific string. Once excited, the waves go on indefinitely unless damped. If a perfect fit of integral de Broglie waves does not occur, the waves will not be in phase, and hence a large number of orbital traversals would lead to complete cancellation of the wave in that orbit. Since the average intensity of the wave, $\overline{\Psi}^2$, is considered a measure of where a particle is located, an orbit with a cancelled wave means that an electron cannot be found in that orbit. An important special case of the quantization rules occurs when n becomes quite large. The energy levels remain discrete until the level where E equals 0. Then the separation of the levels is so fine that the energy states will actually appear to be a continuum. In other words, this is the classical limit. Classical behavior is found beyond this level. An electron raised to the "continuum state" is in fact a free electron, and its energy is non-quantized. Because the attractive energy (given a negative sign in Eq. (8.25)) is the potential energy, when E is greater than or equal to 0, the electron is no longer under the electrostatic control of the nucleus. The energy is dominated by the kinetic motion that can effectively translate the electron away from the nucleus.

What we have described to this point is often called the "old quantum mechanics." It has given way to a more complete formulation using either Heisenberg's matrix mechanics or more commonly Schrödinger's wave equation. We have taken the time to explore the history of the development of the old quantum mechanics because it is an elegant way to see model-making in action and because the "old" models are still valuable as first approximations as we came to know the strange world of quantum mechanics. A significant weakness of the old quantum mechanics was that only systems that are periodic could be treated, and many physical systems of interest are not periodic. In addition, while the total energy of each allowed state in a system could be found, there was nothing in the theory that allowed knowledge of the rate at which the transitions might take place. A corollary problem to this lack of kinetic information was that the theory could not dependably tell which transitions are observed to occur and which are not. From a practical standpoint this theory could only treat one-electron atoms and fails badly for the atoms of biological interest such as carbon, nitrogen, and oxygen. (It treats Na, K, Rb, Cs adequately because they are much like a one-electron atom at least in terms of their valence shells.)

How did these early workers navigate the interface between classical and quantum treatments? This is not just a historical question but is one of enormous practical value because we routinely use both classical and quantum mechanical methods side by side in our research work today (cf. Chapter 26). In 1923 Bohr proposed rules for this interface called the *correspondence principle*. This principle says that the behavior of a physical system as predicted by the quantum theory must correspond to the predictions of classical theory in the limit of large quantum numbers that specify the state of the system. In other words classical mechanics is a special (macroscopic) case of quantum mechanics. Importantly, purely quantum effects such as spin disappear in the correspondence limit.

8.11 Quantum Mechanics Requires the Classical Trajectory Across a Potential Energy Surface to be Replaced by the Wavefunction

Modern quantum mechanics is built on the wave–particle duality, the matter waves of de Broglie, and the quantization of energy. The idea that particles are sharply localized is abandoned and is replaced with the fundamental idea that a particle's position is best treated as a distribution similar to the waves and statistical distributions with which we have become familiar. In this treatment of biophysical chemistry we have focused our interest on the relationship between structure, function, and action and are using the abstraction of the potential energy surface as a central in understanding of this relationship. Basic to that understanding is that *Potential energy surfaces relate position to energy*. Classical mechanics lets us talk about movement across a potential energy surface as a trajectory with each new position and particular momentum determined by the laws of classical mechanics. The path of a particle in the quantum mechanical system cannot be treated as a classical trajectory; instead a new concept, the *wavefunction*, is introduced to replace it. Quantum mechanics is the method for determining the distribution of the wavefunction in a physical system, and an important task is the extraction of the properties of the system from the wavefunction. Quantum biochemistry is the study of the potential energy surface of a system in a way that relates the action and structure of the system to the wavefunction or distribution of position and its energy. There are two methods to find the distribution of position: (1) Heisenberg's matrix mechanics and (2) Schrödinger's wave mechanics. They can be shown to be equal. We will only examine Schrödinger's wave mechanics because it is the most widely used and rests on fairly simple and familiar concepts.

Wave motion can be captured in an equation that is typically derived from Newton's second law in which the wavefunction describes the displacement of a system from equilibrium (see Chapter 30). In quantum mechanics the wavefunction is a mathematical tool or description that depicts the matter wave. No clear physical meaning can be assigned to the wavefunction, but *by axiom* it contains all of the information about the quantum system including (1) the electronic distribution, (2) the average position of the electron, and (3) the energy of the electron in each electronic state. The *Schrödinger* equation is a differential equation whose solutions give the wavefunctions. In most cases the *Schrödinger* equation has boundaries, and like the standing wave example of the Bohr orbits, this limits the number of acceptable wavefunctions. This limited acceptability is the quantization of the system. Once the wavefunction is known it can be operated on by another function called an *operator* which extracts the information of interest from the wavefunction. The value of the extracted property is called the *expectation value*. We will describe several operators shortly. The amplitude (ψ) of the wavefunction relates the probability of finding a particle at that spatial coordinate; the slope (ψ) (i.e., the first derivative) relates the linear momentum of that point of the wavefunction; the curvature (ψ) (i.e., the second derivative) relates the kinetic energy contribution. The expectation values are the sum of each of these local quantities, and these are the observed values of the wavefunction. A distinction between classical wavefunctions and quantum mechanical wavefunctions is that the properties of the classical wavefunction are observables, i.e., directly measured, while the properties of a quantum wave are derived from the process of operating on the wavefunction. Quantum theory includes the observer by directly coupling us to a system and direct observation is impossible.

The wavefunction is a function of the space coordinates and time, and time-dependent equations are important in certain systems (these wavefunctions are labeled Ψ).

$$\Psi = \Psi\ (x, y, z, t) \tag{8.31}$$

For simplicity, we will only be concerned with time-independent expressions, which are labeled ψ:

$$\psi = \psi(x, y, z) \tag{8.32}$$

The electronic distribution or the probability density of finding a particle at a specific point (a finite volume dv) in space is not directly available from ψ. An important interpretation contributed by Born is that ψ^2 characterizes the distribution of the particle in space. This interpretation has analogy to classical electromagnetic wave theory in which the square of the amplitude is the intensity of the radiation. With this interpretation, information about the “position” of the particle is available. Another aspect of our earlier study of probability is that the probability of finding an electron in all space must always be 1. This is called the *normalization condition*; the electron must be somewhere. The electron density is what determines x-ray scattering in a molecule. ψ^2 is a description of the structure of a molecule.

The average position of the electron provides information on the position of the charges in the molecule (i.e., the charge density of the molecule). The charge density assigned to each atom in a molecule along with the structure of the molecule can be used to determine a molecule’s dipole moment. The interaction of the molecular dipole moment with the electrical vector of light leads to many of the spectroscopic effects we use daily in the laboratory to characterize and measure changes in biological molecules. Another aspect of charge distribution is the definition of regions of nucleophilicity or electrophilicity. Our understanding of chemical reactivity is closely tied to the idea that regions of positive charge invite attack by electron-rich domains and vice versa. In these terms we can expect nucleophilic or electrophilic attack in organic reactions. The charge density of a molecule will often change dramatically when an electron is moved into an excited state. Photoexcitation which leads to an excited state with a shift in charge density is one mechanism by which energy transduction occurs in photosynthesis and vision.

The electronic energy of molecules is particularly important in considering movement on a potential energy surface that results in bond formation and breakage. The wavefunction can provide the energy of each electronic state, both the ground state and excited states. By calculating a series of electronic wavefunctions for a set of internuclear separations, the relationship of the energy and the position of the atoms can be found. An analysis of this type can find the stable bond distance in a molecule, determine when occupation of an excited energy state leads to dissociation, and so on (Fig. 8.10).

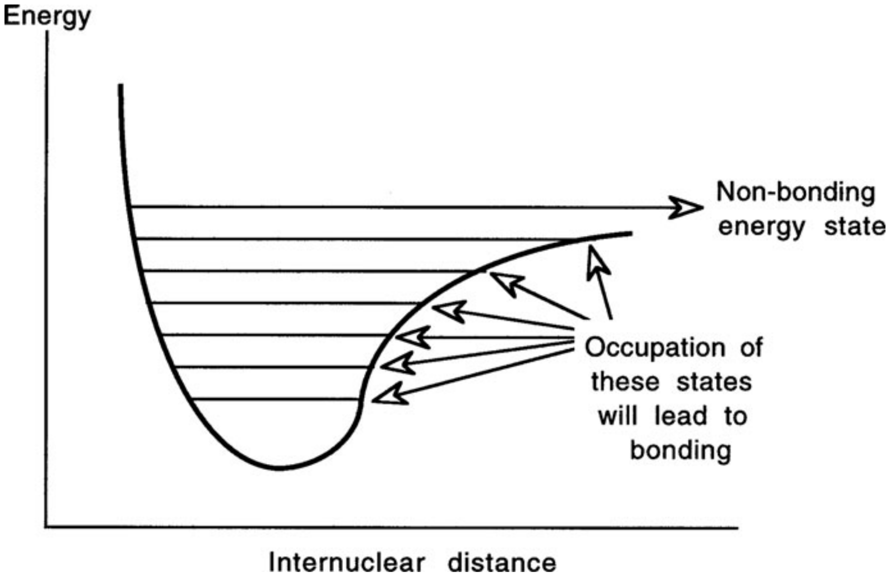


Fig. 8.10 Plots of energy versus internuclear distance can predict bond formation and breaking

8.11.1 The Schrödinger Equation

The Schrödinger equation is often much feared because it looks. . .well. . . fearsome. However, it is nothing more than a mechanical expression for the motion of a particle. In classical mechanics this was not so frightening: the motion, location, and energy of a particle were given by the potential and kinetic energies. The potential energy depends on where in a field (electrostatic, gravity, etc.) a particle finds itself. The kinetic energy depends on the momentum of the particle. The systems that we are familiar with have been

1) linear motion in free space
2) linear motion in a constant field

3) circular motion in a constant field
4) harmonic motion in a parabolic field
5) motion in fields of varying potential (boxes, wells, spherical wells, non-linear potential)

Classically we are able to consider the mechanics of movement in these kind of systems by using energy diagrams (Fig. 8.11). In the classical cases we have been able to treat the position (potential energy) and momentum (kinetic energy) with an arbitrary degree of precision. The Schrödinger equation is no different except that the indeterminacy relation limits our knowledge of the position and momentum of the particle. So we use the wavefunction to describe the position and momentum. The equation says

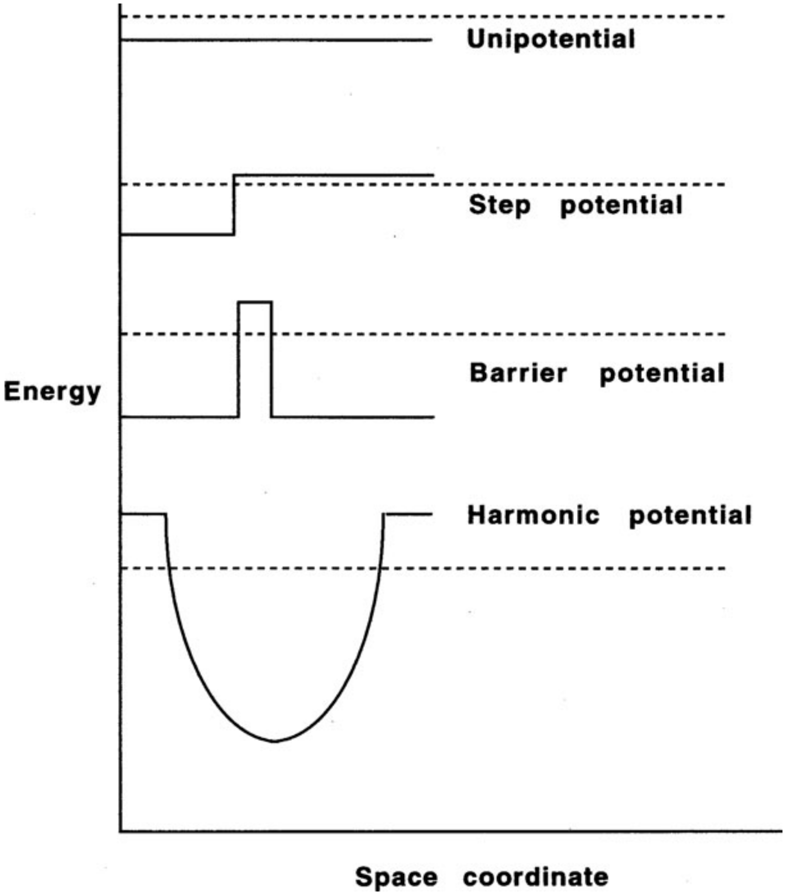


Fig. 8.11 Energy diagrams provide a convenient method to describe the movement in a potential energy field

$$\text{Total energy}\,(\psi) = \text{kinetic energy}\,(\psi) + \text{potential energy}\,(\psi)$$

which, for motion in one dimension, can be written more compactly

$$E\psi = T_x\psi + U(x)\psi \tag{8.33}$$

The kinetic energy of a particle is

$$T_x = \frac{1}{2}mv_x^2 = \frac{p_x^2}{2m} \tag{8.34}$$

The potential energy expression depends on the appropriate expression for the field (Table 8.3).

$$E\psi = \frac{p_x^2}{2m}\psi + U(x)\psi \tag{8.35}$$

Table 8.3 Potential energy functions for consideration in quantum mechanical problems

Particle in a box
$U(x) = 0$ for $0 < x <$ a
$U(x) = \infty$ for $x \le 0$ and $x \ge a$
The Harmonic Oscillator
$U\,(x) = \frac{1}{2}kv^2$
The coulomb potential
$U\,(x) = \frac{q_1q_2}{4\pi\varepsilon_o r}$

In quantum mechanics the expression for momentum must take into account the de Broglie relation. Furthermore the kinetic energy is related to the curvature of ψ at x and so is a second derivative expression:

$$T_x = -\frac{\mathrm{h}^2}{2m}\frac{d^2\psi}{dx^2} \tag{8.36}$$

Then we can write

$$E\psi = -\frac{\mathrm{h}^2}{2m}\frac{d^2\psi}{dx^2} + U(x)\psi \tag{8.37}$$

which, can be simplified by using the Hamiltonian operator, H, which is in general, $H = T + U$:

$$E_n\psi_n = H\psi_n \tag{8.38}$$

The same expression can be used for motion in more than one dimension, which makes the algebra fearsome again but will not at this point enhance our understanding. Any of the standard textbooks on quantum mechanics listed in the end has plenty of these nasty looking (but really quite friendly) equations. The subscript n represents the quantum number of the wavefunction, which is also called at this stage the *eigenfunction*, ψ_n. The Schrödinger equation is a differential equation that can be solved for the values of the energy and the wavefunctions as long as E is constant. The values for E are called the *eigenvalues*, E_n, of the eigenfunctionψ_n.

How do we find the eigenfunctions and eigenvalues for a particular problem? The steps are fairly straightforward though the mathematics can become quite complicated. In general, the process is not difficult. The first point to recognize is that the Hamiltonian expression varies only in the potential energy term because the kinetic energy operator is always the same. The set-up of the problem then becomes What is the potential energy function (path or surface) for this problem? We treat these field functions the same way we are used to treating them. The functions of importance are summarized in Table 8.3. Now comes a key point. What are the boundaries of the problem? We are dealing with continuous functions and they must be single-valued and finite everywhere. For a bound particle the wavefunction must vanish at infinity, in other words the wavefunction vanishes at a potential of $+\infty$. At the nucleus of an atom the wavefunction is undefined because the potential energy becomes $-\infty$.

We now write the Hamiltonian and solve the equation as a differential equation. Realistically many workers at this point will use a computer and numerical methods to find the eigenfunctions and eigenvalues or acceptable approximations to them. Once we have this information we can make chemical arguments based on the characteristic (eigen) energies or wavefunctions that give information about the energy state of a system, the structure, the electronic distribution, etc. (Table 8.4). The following panels summarize in graphic form the solutions to the specific cases in which we have been interested to this point.

Table 8.4 Operators to extract information from the Shrödinger equation for a single particle in Cartesian coordinates

Position	x, y, z
Momentum	$p_x = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial x}$ $p_y = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial y}$ $p_z = \dfrac{h}{2\pi i}\dfrac{\partial}{\partial z}$
Energy	$H = -\dfrac{h^2}{8\pi^2 m}\left(\dfrac{\partial^2}{\partial x^2} + \dfrac{\partial^2}{\partial y^2} + \dfrac{\partial^2}{\partial z^2}\right) + U$

8.12 Solutions to the Time-Independent Schrödinger Theory

8.12.1 Linear Motion – Zero Potential Field

A free classical particle moving in constant-potential space, $U(x) = \text{constant}$, has a trajectory that looks like:

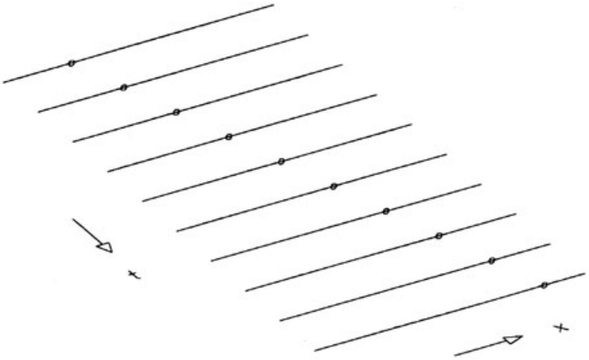


Fig. 8.12 Trajectory of a classical particle in constant-potential space

In terms of its de Broglie wave a particle of defined momentum, $\cos(kx - \omega t)$, looks like:

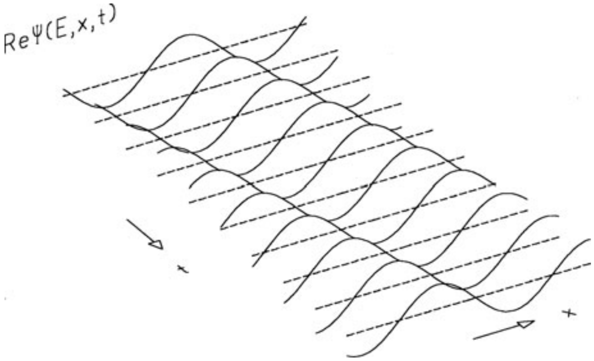


Fig. 8.13 de Broglie wave representation of a particle of constant momentum

Over time, the nodes move in the direction of increasing x but all of the amplitudes are the same and a sense of motion is absent. The wavelength is constant and by the Einstein–de Broglie relation, $E = \mathrm{h}\omega$, this means that the energy and momentum of the particle are perfectly known, $\Delta\mathbf{p} = 0$. The indeterminacy principle thus forces the Δx of the particle to be ∞. The probability density of this pattern over time is constant $\Psi\Psi^* = e^{-i(kx-\omega t)}e^{i(kx-\omega t)} = 1$:

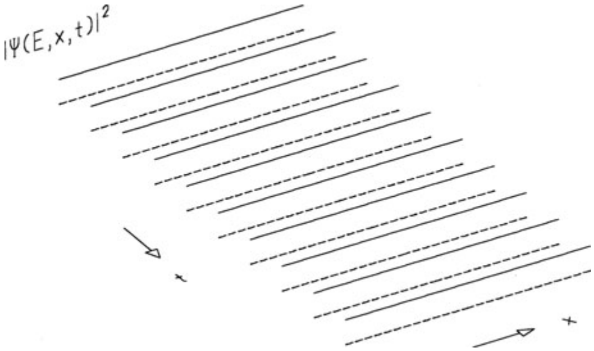


Fig. 8.14 Probability density of the particle/wave represented in Fig. 8.13

This does not leave us with a familiar sense of movement even though the particle is moving with a known velocity through space. To obtain a physically more realistic sense of movement we can create a wave packet by combining a large number of wavefunctions of the same form. Here we show a wave packet created by the in-phase superposition of a Gaussian spectral array and the related probability density of the packet:

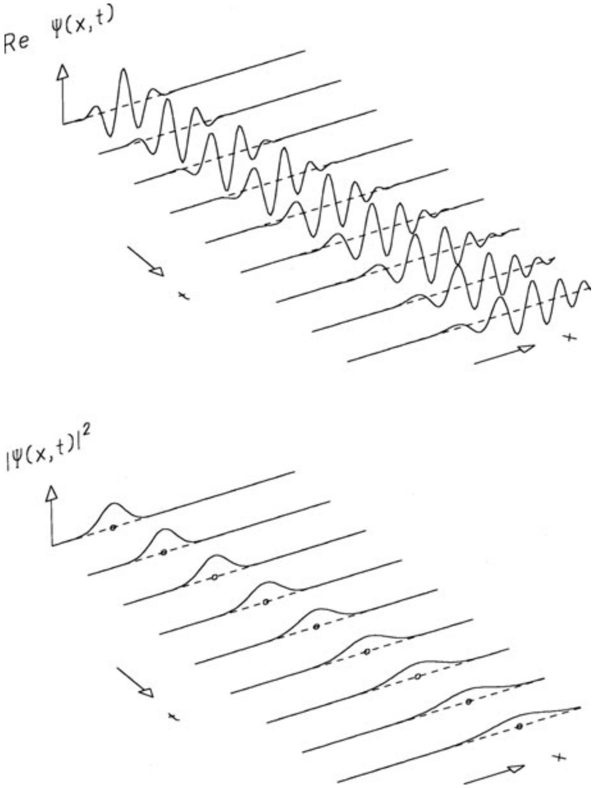


Fig. 8.15 *Top*: apparent motion of a wave packet represents the motion of a particle. *Bottom*: probability density of wave packet shown above

Two points need to be emphasized. The first is that the much more classical movement of the wave packet is easier to interpret because it is closer to the expected (experience of the observer/student) behavior. This has come at the expense of the knowledge of momentum since a large number of wavefunctions each with their own momentum were summed to generate the wave packet. We can identify the position of the wave packet on the x-axis so Δx is no longer ∞, but at the same time we no longer are certain of the momentum, so $\Delta \mathbf{p}$ is no longer equal to 0. In addition, the certainty of the actual position of the particle depends on the width of the wavepacket and our ability to determine its actual wavelength. The second point is that as the particle wave moves it undergoes dispersion. This is a classical property of waves. Both of these effects are due to the indeterminacy principle. In spite of the easier interpretation, the wavepacket method is not frequently used because of the difficult mathematics that results from the superposition of multiple wavefunctions. It is simpler to perform calculations with a single wave number and energy. We will use whichever make the point more easily.

8.12.2 The Step Potential

8.12.2.1 Energy Less Than Step Height

Consider a particle with a particular total energy E_{a} moving in a region of constant potential and then meeting a step-up potential to a second region of constant potential, E_{f}, such that the step height is greater than E_{a}. The classical result will lead to an impulsive force that causes the particle to reverse its direction and travel back the way it came with equal momentum.

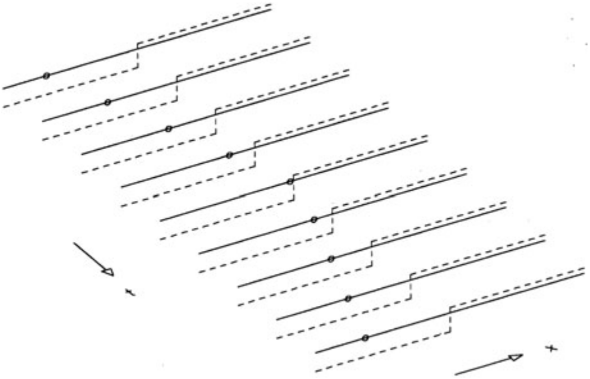


Fig. 8.16 Motion of a particle before and after collision with a barrier of energy larger than the particle

When a wavefunction meets the high potential it cannot suddenly change its direction. This leads to the surprising result that the matter wave penetrates the classically excluded barrier. The penetration of the barrier does not mean that the particle is stored there but rather that it enters and interacts in a classically forbidden zone thus changing its overall behavior. We show the eigenfunction for a particle meeting the step potential with penetration of the barrier:

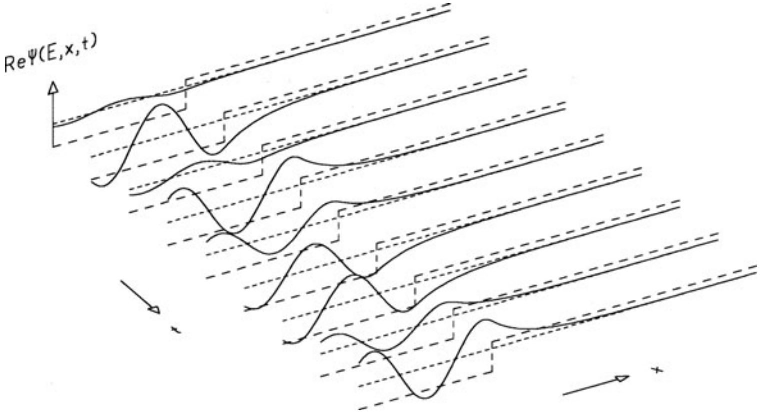


Fig. 8.17 Penetration of the step potential by a wave

Here is a wave packet approaching and decomposing with penetration of the classical forbidden zone and then reflecting back changed. The sawtooth patterns seen at the barriers are due to interference as the reflected wave interacts with the incident wave:

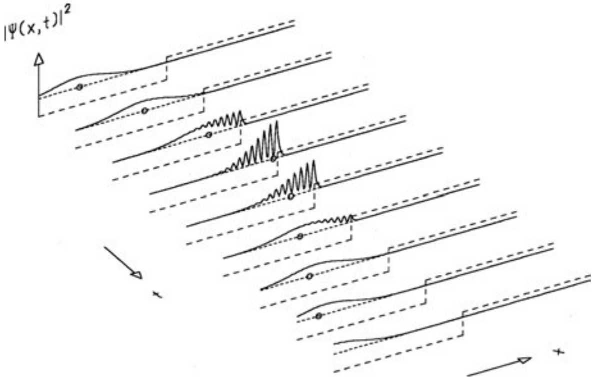


Fig. 8.18 Penetration of the step potential by a wave packet

8.12.3 The Barrier Potential

Consider a barrier potential at $U(x)$ and height U_0. In a classical system if the particle has an energy $> U_0$, it will be reflected with a probability of 1. If it has an energy < U_0, it will pass through the barrier with a probability of 1 and be transmitted with a reduced momentum. The quantum mechanical result is quite different. Even if the particle has an energy greater than the barrier, there is a finite probability that the particle will be reflected except for certain values of E. Alternatively, if the energy is somewhat smaller than the barrier, there is finite probability that the particle will be transmitted through the barrier into the forbidden zone. Here are three barrier heights with the mixed reflection/barrier penetration shown (top to bottom, the barrier potential increases).

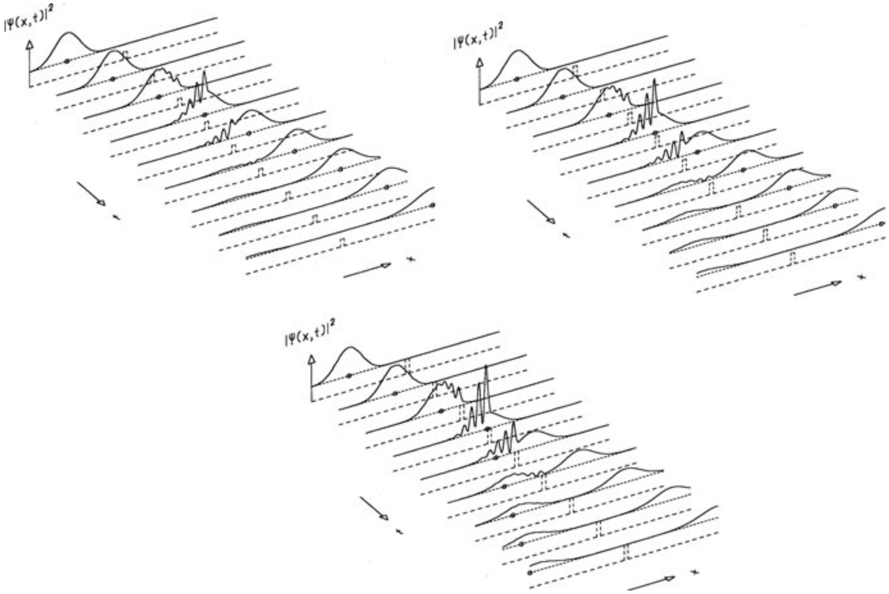


Fig. 8.19 *Top*: barrier interaction by a particle and its wave representation when the energy of the particle is greater than the barrier. *Middle*: barrier interaction when the particle energy is just greater than that of the barrier. *Bottom*: barrier interaction with tunneling when the particle energy is just below the barrier potential

The remarkable aspect of this *tunneling* through the barrier is that the tunneling is wholly a quality of a wave yet it is a particle that can be detected in the region beyond the barrier. This tunneling phenomenon plays an important role in a wide variety of processes including the emission of a particle from radioactive nuclei, the inversion of the ammonia molecule, the high-speed switching action of the tunnel diode in electronics, the transfer of electrons in photosynthesis and electron transport, and the anomalously high conductivity of protons in water. The transmission and reflection coefficients depend on the relationship of the height of the barrier and the barrier thickness when E is a reasonable fraction of U_0. Compare the result of varying the barrier height (*top* = *middle*) and width (*top* = *bottom*) in Fig. 8.20.

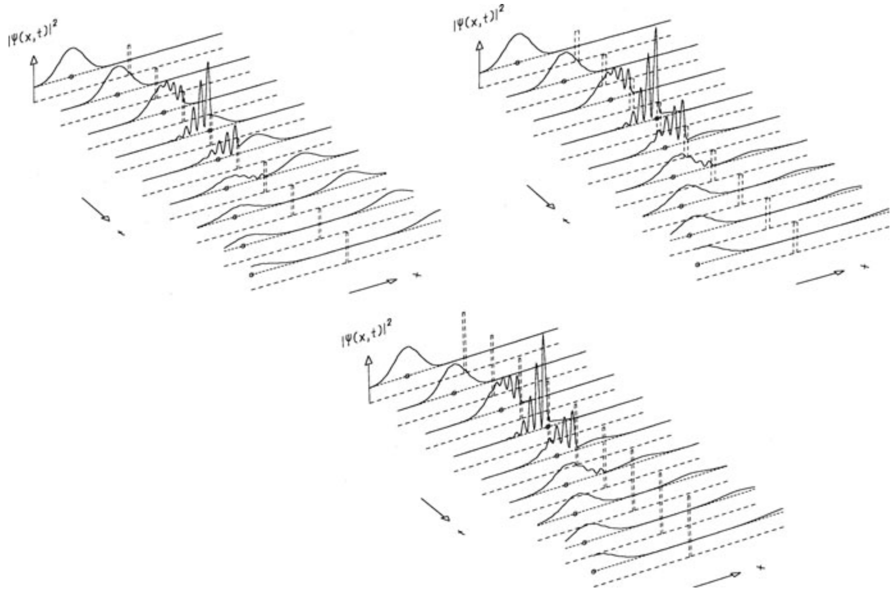


Fig. 8.20 Transmission and reflection of a wave/particle with varying barrier heights and widths. *Top* and *middle* barrier heights are equivalent. *Top* and *bottom* barrier widths are equivalent

8.12.4 The Square Well Potential

Though we have seen a variety of interesting quantum mechanical effects in the freely moving particles we have not seen any quantization. Quantization occurs only when a particle is bound to localized regions of space. Local boundaries for a particle provide the picture of atomic and molecular structure. The simplest of these potentials is the square well potential or the particle-in-a-box problem. This potential has extensive application in biological systems as a useful abstraction. It can be used to approximate the behavior of delocalized electrons bounded in potential energy barriers which is reasonably close to the way we have been treating structure and the potential energy surface.

The particle-in-a-box is solved by treating the space inside the box (or cube or sphere) as a region of constant potential with potential energy boundaries of $+\infty$. Applying the boundary conditions, as we mentioned before, means that at the boundaries the wavefunction must take a value of 0. Inside the box the wavefunction will be a trapped de Broglie wave.

It is appropriate to treat the possible wavefunctions as if they are standing waves on a piece of string of fixed length. The wavefunctions allowed will be those that have an amplitude that vanishes at the edges of the box where $U = \infty$. Like a string in a standing wave system, the higher the energy of the system, the more nodes of the wave will be seen:

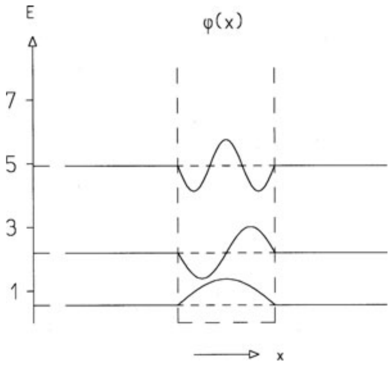


Fig. 8.21 Examples of the wavefunctions allowed in the particle-in-a-box formulation

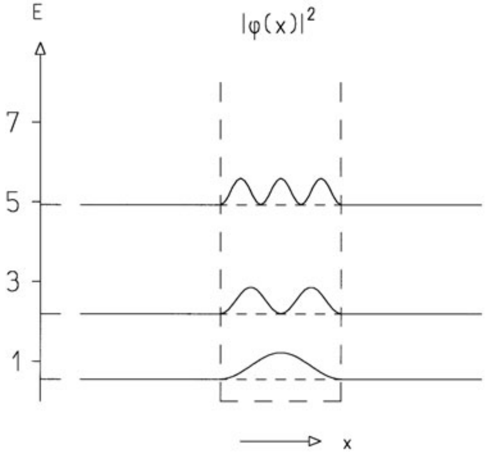


Fig. 8.22 Higher energies are associated with greater node counts in the particle-in-a box

The relationship of the energy to the quantum number is exponential and so the spacing of the allowed energy levels increases as a series.

The length of the box is akin to determining the length of the vibrating string, and so a longer dimension will have longer wavelengths and lower energies associated with the quantum numbers.

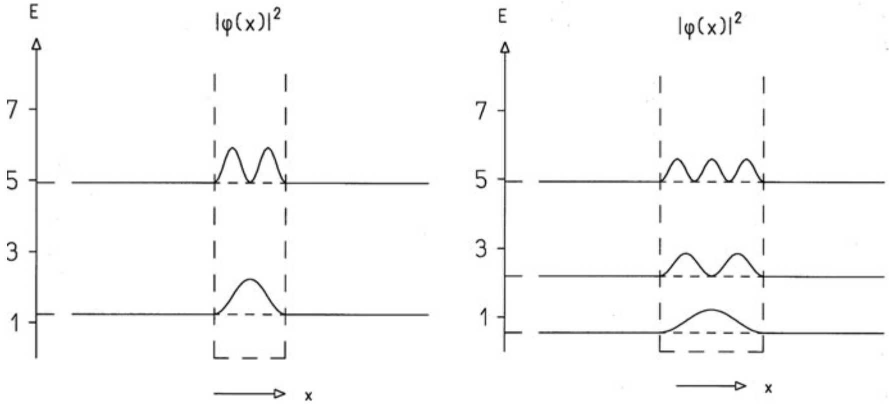


Fig. 8.23 Four examples of varying widths of the particle-in-a-box

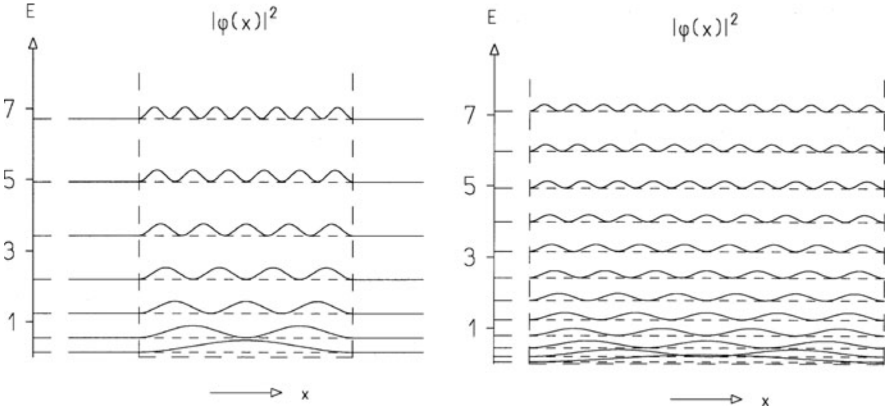


Fig. 8.23 (continued)

This type of analysis is applied qualitatively to molecules with delocalized π electrons such as aromatic hydrocarbons and aromatic heterocyclics (tyrosine, pyrroloquinoline quinone), long chain polyenes (retinal and β-carotinoids), and porphyrins (heme, chlorophyll). An electron occupying a lower energy level can be excited to a higher energy level if it absorbs the correct energy photon. The energies for these transitions are carried by photons with wavelengths in the visible and ultraviolet range; this is the link between UV–visible molecular spectroscopy and electronic structure that we use in biophysical studies. We will explore several examples of this in the sections on spectroscopy.

We have described the ideal case of a box formed with walls of infinite height. Generally this is not the case, and penetration of the barrier potential with the expectation of tunneling probabilities must be considered for each wavefunction. Here a variety of well widths and depths are shown with the corresponding wavefunctions. Varying penetration of the step potentials can be appreciated.

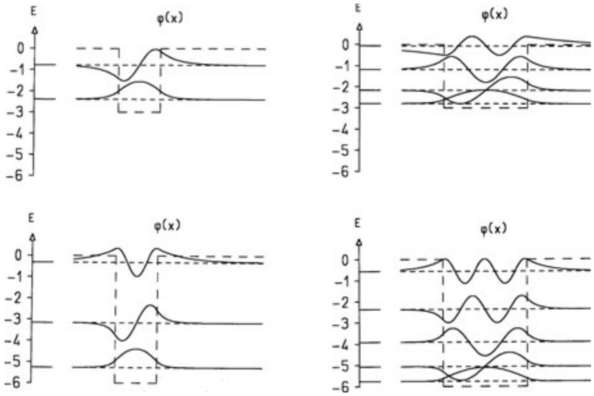


Fig. 8.24 Varied well depths and widths in the square wave potential

The combination of chemical, electrochemical, or photo-excitation with tunneling is an area of active study in a wide variety of areas such as photosynthesis, energy transduction, and enzymology.

8.12.5 The Harmonic Oscillator

Harmonic oscillations occur when a body experiences a restoring force that obeys Hooke's law. This implies a parabolic potential. The maximum potential energy occurs at the turning points, and the maximum kinetic energy occurs at the point of zero displacement. In a classical oscillator, the particle will most likely be found at the turning points because it moves most slowly in that vicinity. The potential energy curve and probability density of a classical oscillator are shown along with the probability density for a quantum mechanical oscillator in the lowest energy state. The results are strikingly different.

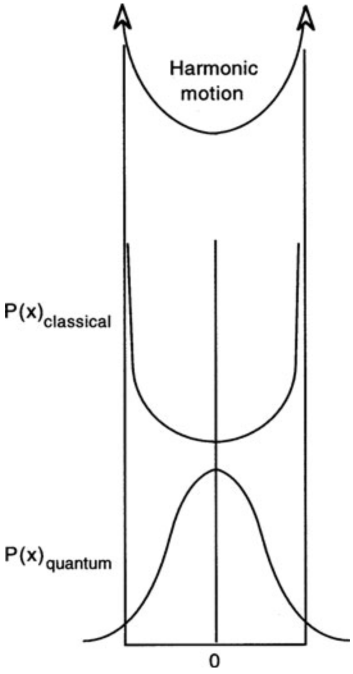


Fig. 8.25 Classical and quantum probabilities for harmonic motion

Applying the Schrödinger equation to a system experiencing a parabolic potential gives the following picture of wavefunctions that fit into the potential energy well.

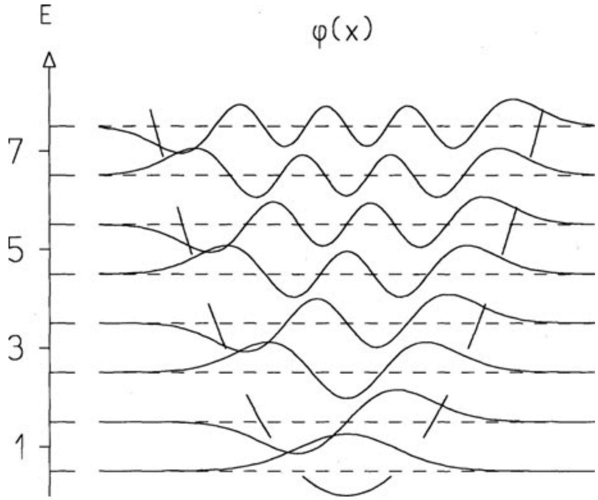


Fig. 8.26 Wavefunction associated with a parabolic potential

The harmonic oscillator is a model of vibrational modes of energy and thus has important application to our understanding of molecular movement. These vibrations are the bond oscillations around the equilibrium point and are important in studies involving bond stretching and bond bending. The energy of the oscillator is quantized and has evenly spaced energy levels. There is a zero-point energy that is an important result of the indeterminacy principle. There can be no occupation of the position of zero displacement because at this point $\Delta x = 0$ and $\Delta \mathbf{p} = 0$, and the principle of indeterminacy would be violated. This provides us with an important result: A system at the lowest energy level still possesses energy. This is the quantum basis of the third law of thermodynamics. The photons necessary to cause transitions between the energy levels of a harmonic oscillator fall in the infrared region of the electromagnetic spectra. This is the underlying basis for infrared spectroscopy and its application to structure determination in biochemical studies.

The quantum prediction of the probability density is quite different from that of the classical oscillator. The correspondence principle is obeyed, however, as we examine wavefunctions of higher *n*. At these higher numbers, as the classical limit is approached, the wavefunctions progressively begin to resemble the classical probability curve:

8.12.6 Rotational and Angular Motion

Other potential energy fields can be explored that provide a quantum mechanical solution to rotational motion such as treatment of a particle-on-a-string or the

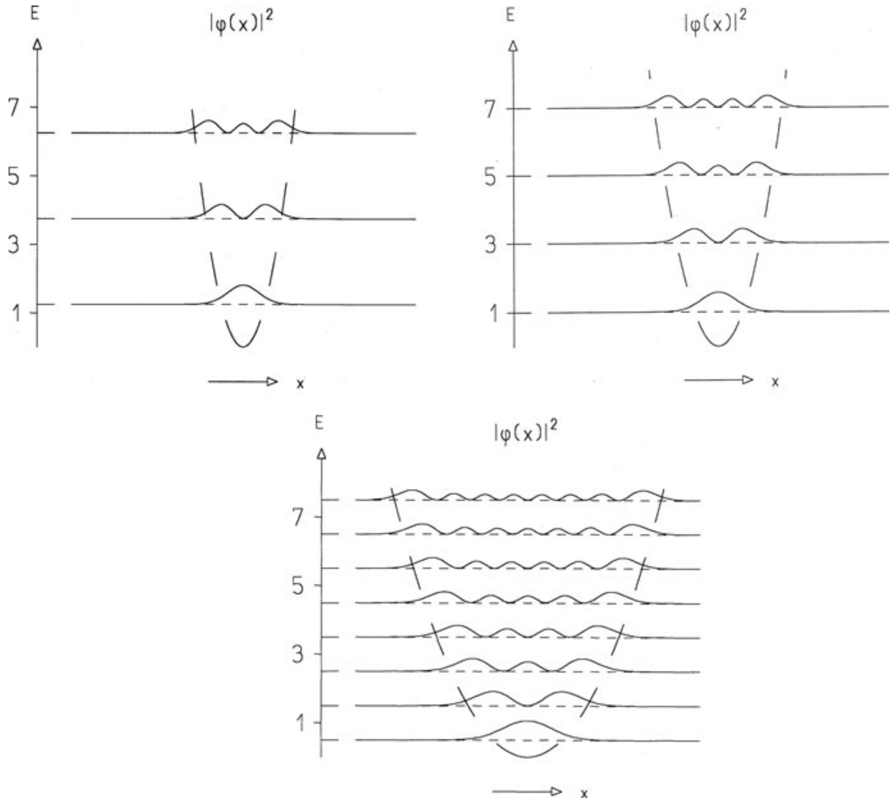


Fig. 8.27 Parabolic potentials of varying widths

particle-on-a-sphere. These solutions are important in understanding rotational energies and can be studied with microwave spectra; but their practical applicability is restricted in general to dilute gases. We will not discuss them in any detail here, but the interested student can find them thoroughly discussed in references given at the end of the chapter. One aspect of rotational motion that must be appreciated is the angular momentum that a rotating system possesses. Classically angular momentum, which is a vector whose direction is determined by the right-hand rule, may have any magnitude and orientation. However, the magnitude of the vector in quantum mechanics is quantized and limited to values given by $l = \left\{ j\,(j+1)^{1/2} \right\}$h where j is a non-negative integer or in the case of electrons (and all fermions) a half integer. Quantum mechanics limits our knowledge of the discrete value of the components of angular momentum to just one. The axis that is favored is defined as the z axis and can be determined by applying an external electrical or magnetic field. Because the x and y components are indeterminate, the form of the quantized vectors can be rotated around the z axis and the vector may lie anywhere within the cone. This angular momentum becomes important in building an atomic model.

8.13 Building the Atomic Model – One-Electron Atoms

We now arrive at a very important potential field, the *coulomb field*. By solving the Schrödinger equation for the motion of a single electron in this field we will be calculating the hydrogenic wavefunctions. Our model is an electron with its associated matter wave occupying quantized distributions in space bound by the coulomb field of the nucleus. The total energy will be given by the principal quantum number, n. The electron will have momentum around the nucleus; thus it will have angular momentum, $l = 0, 1, 2, ..., (n-1)$ .. Because the charged electron is apparently circulating it has a magnetic field associated with this electrical current, and the magnetic quantum number, m_l, can be any integer between ±l. Finally electrons are fermions with $j = \frac{1}{2}$ or spin $= \frac{1}{2}$ and have an intrinsic magnetic property, $m_s = \pm\frac{1}{2}$.

The mathematics looks like this

$$E_n = -\frac{2\pi^2 \mu e^4}{(4\pi \varepsilon_o)^2 \; h^2 n^2} \tag{8.39}$$

The probability densities in one dimension look like this:

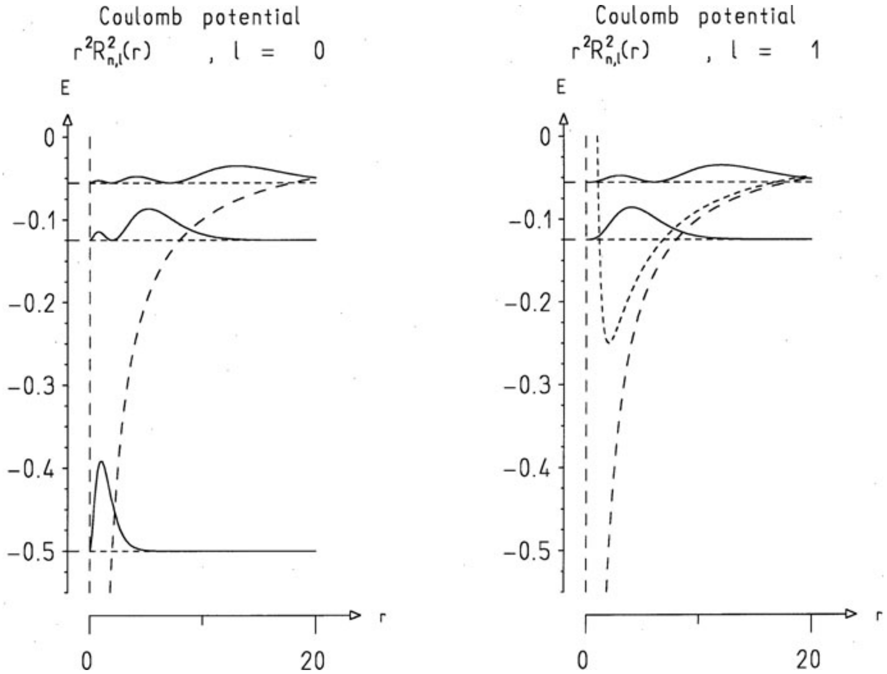


Fig. 8.28 Coulomb potentials and associated probability densities in one dimension

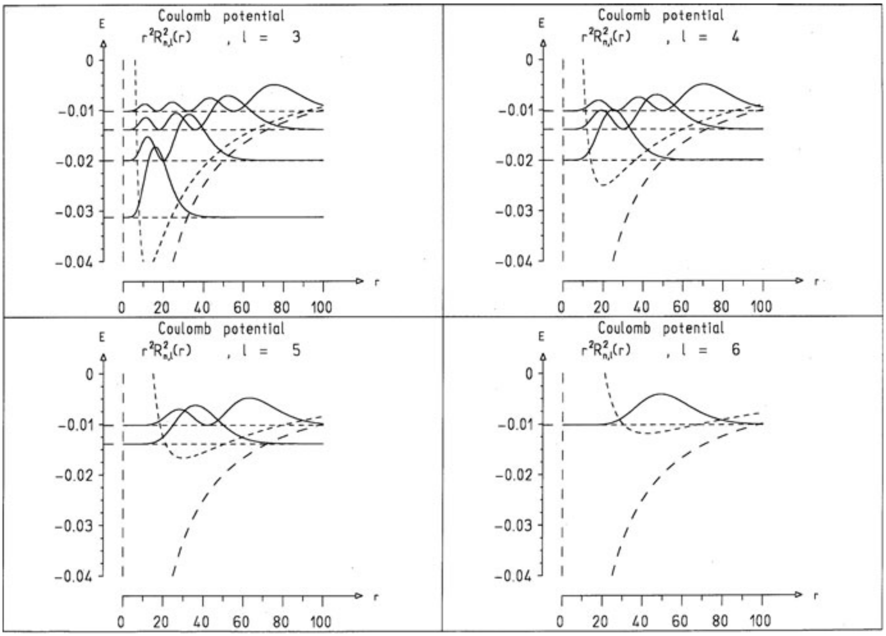


Fig. 8.28 (continued)

In three dimensions, as functions of (*n*, *l*, *m*), they look like this:

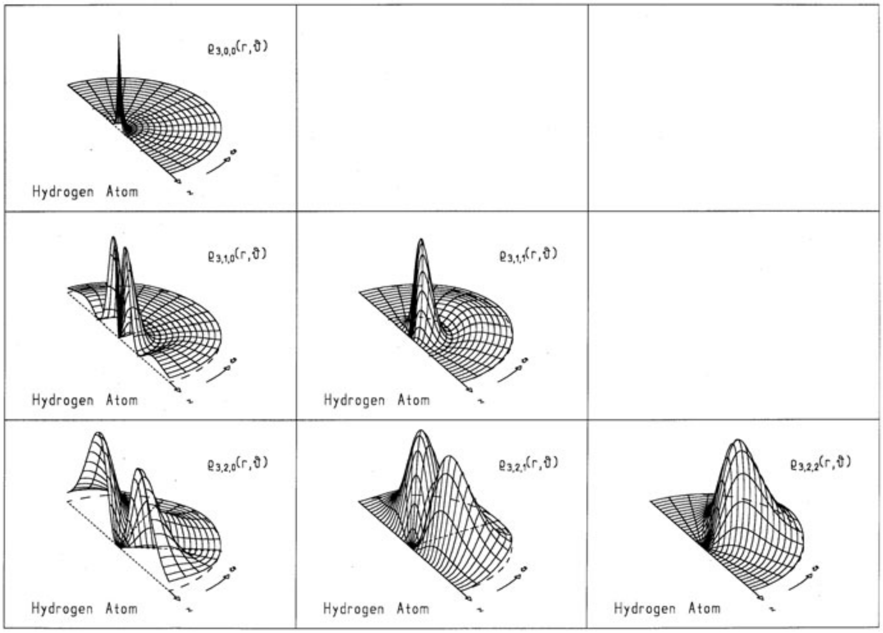

$\varrho_{3,0,0}(r,\vartheta)$
Hydrogen Atom
$\varrho_{3,1,0}(r,\vartheta)$
Hydrogen Atom
$\varrho_{3,1,1}(r,\vartheta)$
Hydrogen Atom
$\varrho_{3,2,0}(r,\vartheta)$
Hydrogen Atom
$\varrho_{3,2,1}(r,\vartheta)$
Hydrogen Atom
$\varrho_{3,2,2}(r,\vartheta)$
Hydrogen Atom

Fig. 8.29 Coulomb potentials and associated probability densities in three dimensions

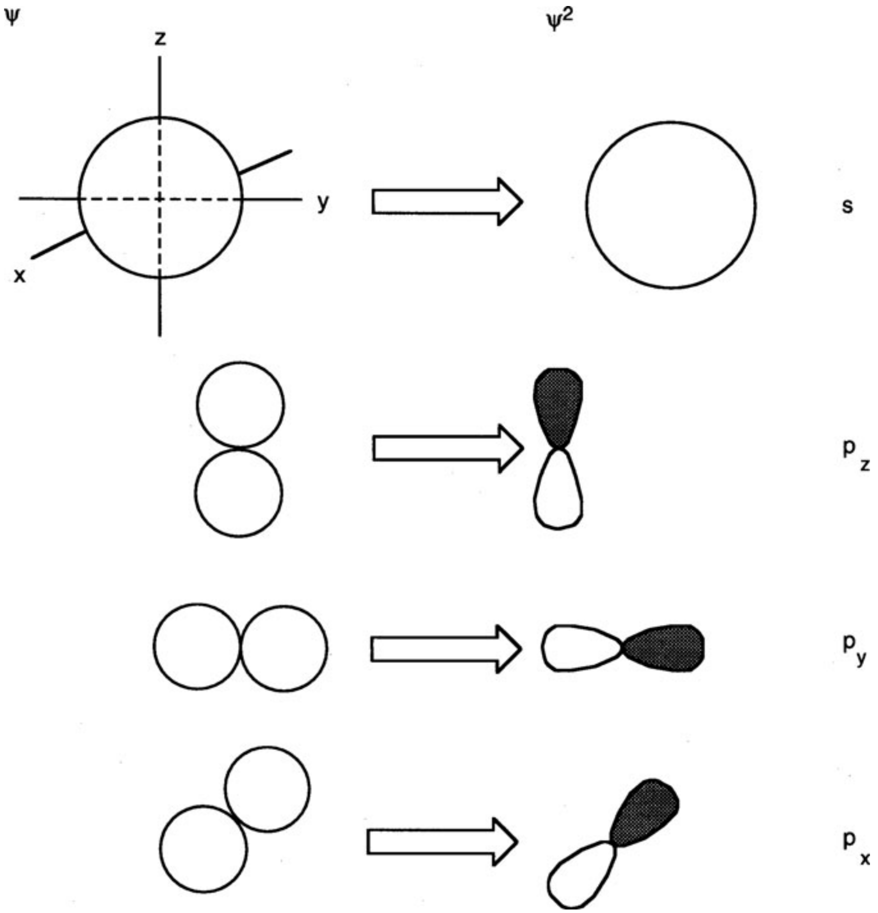


Fig. 8.29 (continued)

These plots are familiar: they are the shapes of orbitals for a one-electron atom. With $l = 0$ the orbital is spherical, the *s* orbital. When $l = 2$ the *p* orbitals are seen and when $l = 3$ the *d* orbitals are rendered. These plots are representations of the electron density, and the information is derived from operating on the hydrogenic wavefunctions. In terms of atomic and molecular structure these are the only exact solutions to the Schrödinger equation. However, we use a variety of effective approximations such as the linear combination of atomic orbital-molecular orbital method (LCAO-MO) to provide quantitative data that are generally accurate within experimental error. We will explore these areas in the next chapter as they pertain to building useful abstractions to biochemical state space.

8.14 Building the Atomic Model – Multi-electron Atoms

We have spent time on the behavior of particles such as electrons and protons in general space and bounded space because it underlies the whole of chemistry. All of the preceding arguments have exact solutions of the Schrödinger equation. The computational aspects of the next level of complexity, which include multi-electron atoms and molecules, are not solvable except by approximation. But to understand biophysical processes at a chemical level requires some operational system that we can use to systemize molecules of interest. Conceptually this is not as hard as it might seem though the details and mathematics can be quite daunting.

8.14.1 Fermions and Bosons

It is a unique property of quantum mechanics that particles in the Universe can be broken into two fundamental classes, *fermions* and *bosons*. These particles can be differentiated by their spin, which is an integral property of a particle like its charge. Spin is a wholly quantum mechanical property (i.e., it disappears in the classical limit). Fermions are the particles from which matter is formed, and bosons are the particles that have been identified as the conveyors of force in the Universe. For example, electrons and protons are fermions. Photons which convey the electromagnetic force are bosons. Fermions have a spin of $\frac{1}{2}$, while bosons have spin of 1. Fermions satisfy Fermi–Dirac statistics and bosons satisfy Bose–Einstein statistics. A fundamental principle of quantum mechanics is the Pauli principle, which leads to the Pauli exclusion principle forbidding two identical fermions from occupying the same quantum mechanical state. (The Pauli principle dictates that bosons may occupy a given quantum state in unlimited numbers, thus an intense, monochromatic, coherent beam of light can be prepared. This is the principle of the maser or laser.) A maximum of two electrons may occupy a given quantum mechanical state (orbital) in a molecule if one is spin up ($+\frac{1}{2}$, called an α electron) and the other is spin down ($-\frac{1}{2}$, called a β electron). The exclusion principle underlies the principles of chemistry and therefore life because it governs the build-up of atomic and valence structure in the elements (the *aufbau* principle). The exclusion principle is the reason for the electron pair as a basis of chemical bonding. The Pauli exclusion principle is also the reason that molecules cannot be forced too closely together. The consequence of pressing atoms too closely is that more than two electrons would necessarily occupy the same orbital. This is so energetically unfavorable that the repulsive forces prevent this close approach. We will quantify these Pauli exclusion forces as the repulsive forces in our discussion of intermolecular forces. Without the Pauli exclusion, the Universe could be a homogeneous mass. It is important to recognize that it is not only a coulomb repulsion that prevents electrons of the same spin from approaching one another. The most important reason is the Pauli exclusion.

The occupancy of orbitals by electrons of different spin is entitled the *multiplicity*. The multiplicity of an atom is given by a number 2S +1. A system with an even number of electrons is usually a closed shell in the ground state and is described as in the singlet state (multiplicity = 1). A system with an odd number of electrons (there will be one unpaired electron) is a free radical and usually is a doublet (multiplicity = 2). The first excited state of a system usually is a triplet (multiplicity = 3). Diatomic oxygen is a triplet state because the two highest occupied molecular orbitals (HOMOs) are degenerate and occupied by one α electron each.

8.14.2 Self-Consistent Field Theory Finds Approximate Wavefunctions for Multi-electron Atoms

The method for finding the approximate wavefunctions of a multi-electron atom was first proposed by Douglas Hartree. It was modified by Vladimir Fock, and today is called *Hartree–Fock theory* or the *self-consistent field method* (SCF). The method is a compromise between:

1) A first approximation that considers the coulomb interaction between each electron and between the nucleus and each electron. Depending on the relative position of each electron to another electron and to the nucleus, the forces will vary considerably.
2) The requirement that the Schrödinger equation be solvable. This requires that each atomic electron be treated independently. Then with each electron represented by a single time-independent equation, a series of equations can be solved in a fairly straightforward way.

The problem with these two requirements is that they are mutually exclusive. The coulomb potentials between electrons cannot be considered and still meet the condition that equations can be written in which electrons are treated as independent. The compromise is to consider each electron independently moving in a potential field comprised of the contributions of the other electrons. This is done by considering the other electrons and nucleus as a spherically symmetrical net potential $U(r)$ where r is the radial distance from the nucleus. The net field is the average potential contributed by: the attractive interaction between the nucleus and electron, and the negative interaction between the electron and its Z – 1 compatriots. This idea of reducing individual interactions to a net continuum is a method we will see used again in this volume both in the Debye–Hückle theory and in the Kirkwood treatment of solvent. Once this potential is estimated, the Schrödinger equation is solved for the orbital of the electron. Each electron is treated this way with a new set of orbitals produced. At the end of the first cycle, the centrosymmetric net field is recalculated using the new (improved) orbitals of each electron. This process continues until there is no change in the orbitals between cycles. At this point the orbitals are said to be self-consistent and the process is complete. The SCF technique can be applied to molecules through the LCAO method and the molecular orbitals are processed until they are self-consistent.

We now have in hand the basic mechanics of atomic structure and molecular interaction. In the coming chapters, this foundation will be used to explore biological structure and function as well as some of the methodologies used to elucidate the biological state space.

Further Reading

History

Cassidy D.C. (1992) Heisenberg, uncertainty and the quantum revolution, *Sci. Am.*, **266, 5**: 106–112.

Hoffman B. (1947) *The Strange Story of the Quantum*. Dover Books, New York.

Snow C.P. (1981) *The Physicists: A Generation that Changed the World.* Little, Brown and Co., Boston.

Textbook Chapters Treating Quantum Mechanics

Feynman R.P., Leighton R.B., and Sands M. (1963) *The Feynman Lectures on Physics,* Volume 3. Addison-Wesley, Reading, MA.

Tinocco I., Sauer K., Wang J.C., and Puglisi I. (2001) *Physical Chemistry (Principles and Applications in the Biological Sciences),* 4th edition. Prentice-Hall, Englewood Cliffs, NJ.

Textbooks on Quantum Mechanics

Atkins P.W. (1991) *Quanta: A Handbook of Concepts*, 2nd edition. Oxford University Press, Oxford. (Like an illustrated glossary. Fun to read and useful.)

Atkins P.W. and Friedman R.S. (2005) *Molecular Quantum Mechanics*, 4th edition. Oxford University Press, New York. (For the chemically oriented reader, this is the definitive textbook.)

Brandt S. and Dahmen H.D. (1995) *The Picture Book of Quantum Mechanics*, 2nd edition. Springer-Verlag, New York. (Mostly physics and not at an elementary level but the illustrations are very useful and quite beautiful.)

Brandt S. and Dahmen H.D. (1995) *Quantum Mechanics on the Macintosh* (Also available on the PC), 2nd edition. Springer-Verlag, New York. (This is the program used in the book listed above to generate the graphics. I used this program to generate the illustrations for Section 8.12. The program is interactive and acts like a laboratory in which the user can gain a substantial practical feel for quantum mechanical behavior. I highly recommend its use in the classroom, laboratory or even as a “video distraction” for anyone interested in developing more than an arms length feeling for quantum mechanical systems.)

Brandt S., Dahmen H.D., and Stoh T. (2003) *Interactive Quantum Mechanics.* Springer-Verlag, New York. (This is the new version of the earlier books by the above authors.)

Eisberg R. and Resnick R (1985) *Quantum Physics of Atoms, Molecules, Solids, Nuclei and Particles*, 2nd edition. Wiley, New York. (A physics text. But this text is a wonderfully lucid discussion of quantum physics for the student ready to go beyond the introductory treatments.)

Hanna M.W. (1981) *Quantum Mechanics in Chemistry*. Benjamin/Cummings Publishing, Menlo Park, CA.

Interesting Papers

Boeyens J.C.A. (1995) Understanding electron spin, *J. Chem. Educ.*, **72:**412–415.

Corkern W.H. and Holmes L.H. Jr. (1991) Why there's frost on the pumpkin, *J. Chem. Educ.*, **68:**825. (Application of black-body radiation physics to a question dealing with phase transitions.)

Dence J.B. (1983) Note on a Simple Derivation of Planck's Formula from Special Relativity, *J. Chem. Educ.*, **60**:645–646. (The derivation of Planck's important formula from a completely different viewpoint than the historical one we've taken. A good start into relativity.)

Einstein A. (1905) On a heuristic point of view concerning the production and transformation of light, *Annalen der Physik*, **17:**132–148. (Einstein's papers are always worth a careful read. They are surprisingly accessible.)

Einstein A. (1906) On the theory of light production and light absorption, *Annalen der Physik*, **20:**196–206.

Einstein A. (1907) Planck's theory of radiation and the theory of specific heat, *Annalen der Physik*, **22:**180–190.

Englert B., Scully M.O., and Walther H. (1994) The duality in matter and light, *Sci. Am.*, **271, 6**:86–92. (An accessible introduction to complementary properties.)

Volkamer K. and Lerom M.W. (1992) More about the particle-in-a-box system, *J. Chem. Educ.*, **69**:100–107.

Quantum Mechanics in Biology

Devault D. (1980) Quantum mechanical tunneling in biological systems, *Quart. Rev. Biophys.*, **13**:387–564. (Detailed application of this important aspect of quantum mechanics to biological systems.)

Problem Sets

1. The Earth's Sun has a surface temperature of 5800 K. What is its color?
2. Determine the effect of altering the temperature of a star (or going to another star system on the distribution of light and its effect on photosynthesis given the absorption spectrum of chlorophyll versus rhodopsin versus retinol.
3. (a) Show that $e^{hv/kT} \rightarrow 1 + hv/kT$ for the condition $hv/kt \rightarrow 0$.
   (b) Show that $e^{hv/kT} \rightarrow \infty$ for the condition $hv/kt \rightarrow \infty$.
4. Show that (a) the Rayleigh–Jeans law is a special case of Planck distribution law for the blackbody spectrum. Show also that (b) Stephan's law and (c) the Wein displacement law can be derived from Planck's distribution law.
5. At what mass and length would a pendulum have to be constructed to reach the level of demonstrating a discontinuous energy loss?

6. Which biologically important entities should be treated as quantized (Hint: consider the size and speed of the entity and apply the deBroglie relationship and then calculate $\Delta E/E$).
7. If the Sun had a temperature of 7800 K, what would be the ideal length for an isoprene molecule to shield against the peak light spectra from the star?
8. There is a very striking and significant conclusion that must be drawn about the wave theory of light from the behavior of the photoelectric effect. What is it?
9. Explain the cause of Fraunhöfer lines.

Chapter 9
Chemical Principles

Contents

9.1 Knowing the Distribution of Electrons in Molecules Is Essential for Understanding Chemical Structure and Behavior

We will now explore the principal methods for the description and calculation of the electronic distribution in molecules. The importance of having knowledge of a biological molecule's electron distribution cannot be overstated. The electronic distribution in a molecule is responsible for its chemical properties. As we will see, the formation of chemical bonds with the consequent bond strengths, lengths, force

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_9, © Springer Science+Business Media, LLC 2010

constants, and angles defines the shape and reactivity of molecules. The polarity of bonds determines charge distribution and give rise to the dipole moment of a molecule. The interaction of the electronic distribution with photons determines the optical character and spectral properties of molecules and the source of color and photochemistry. Many of the interactional forces on a large scale, including dipole and quadruple interactions, dispersion forces, and hydrogen bonding, all depend on the electronic structure of the molecules. Our appreciation of the chemistry and chemical biology of living organisms is built on an understanding of the electronic structure.

9.2 The Nature of Chemical Interactions

Except in the case of nuclear chemistry and Pauli exclusion, all interactions between molecules are electronic in nature. For the chemistry student, the most commonly considered reactions between molecules are those where electrons are actually shared, leading to covalent bonding interactions. The most accurate view of chemical interactions is a continuum of electronic charge distribution in which one extreme is the nonpolar covalent bond and the other extreme is the ionic bond. The nonpolar covalent bond is characterized by a symmetrically shared distribution of electronic charge that draws the atomic nuclei together, whereas in the ionic bond there are oppositely charged atomic centers that do not share electronic charge but rather are influenced by the electric field between the centers of charge. The familiar ionic bond that underlies salt crystal interactions is the “purest” form of electrostatic interactions that also include van der Waals interactions on the same electrostatic spectrum. In the following sections, we will examine the principles of interaction of the non-covalent and the covalent bonds.

9.3 Electrostatic Forces Describe the Interactions from Salt Crystals to van der Waals Attraction

9.3.1 Ion–Ion Interactions

Ion–ion interactions are common in a pure electrolyte such as KCl. Two charged particles exert an interactive force on one another. The potential energy of interaction, *U*, is derived in Appendix J and is given here:

$$U_{i\text{-}i} = \frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon r} \tag{9.1}$$

Because the potential energy falls off linearly with respect to distance between the two particles, these forces are substantial and dominate at long distances.

9.3.2 Ion–Dipole Interactions

Ion–dipole interactions occur when one of the molecules is an ion and the other a *dipole*. Dipoles carry no net charge yet have a permanent charge separation due to the nature of the electronic distribution within the molecule itself. The concept of the molecular dipole can be illustrated by considering the structure of water and carbon dioxide as in Fig. 9.1. Oxygen is more electronegative than either hydrogen or carbon, and the shared electrons in these molecules will spend more time near the oxygen atoms. The result of this preference for the oxygen atom by the negatively charged electrons will result in a partial negative charge on the oxygen atoms, whereas the companion carbon or hydrogen atoms will carry a partial positive charge. This charge asymmetry in the molecule is given a magnitude and direction by calculating the *dipole moment*, μ. The dipole moment is given in *debye* and is found by multiplying the separated charge times the distance of separation. The existence of a dipole moment requires an asymmetric molecule. As Fig. 9.1 shows, H_2O has a dipole moment of 1.85 debye and CO_2 has no dipole moment, because H_2O is asymmetric and CO_2 is symmetric. In CO_2, the electric vectors derived from the partial charges cancel one another, and no dipole moment is measured. On the other hand, the H_2O molecule is bent and the electric vectors are added, resulting in the formation of an *electric dipole*. The electric dipole is treated like any other charge, although the orientation of the dipoles ($\cos\theta$) must be taken into account in calculating the ion–dipole force (Fig. 9.2). Ion–dipole interactions will be discussed at greater length in Chapter 15. $U_{i\text{–}d}$ is calculated as

$$U_{i\text{-}d} = \frac{q_1\mu_2\cos\theta}{4\pi\varepsilon_0\varepsilon r^2} \tag{9.2}$$

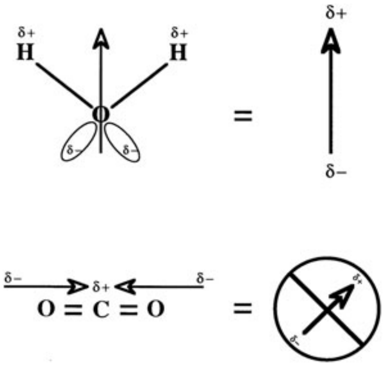


Fig. 9.1 Molecular models of carbon dioxide and water for consideration of the formation of molecular dipoles

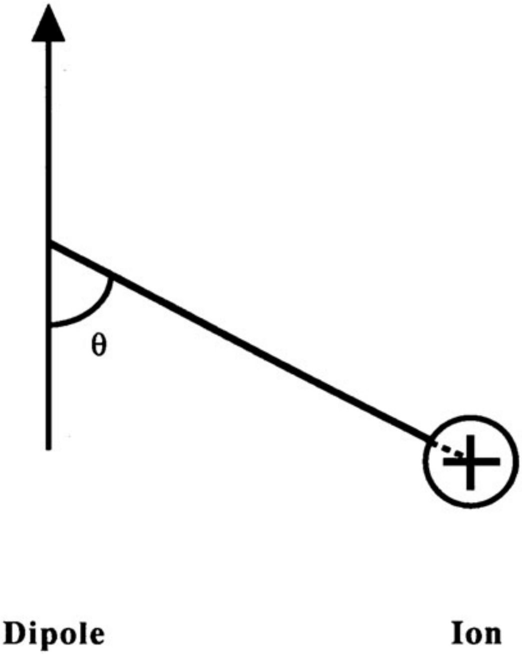


Fig. 9.2 The orientation of the dipoles ($\cos\theta$) must be taken into account in calculating the ion–dipole force

9.3.3 Ion-Induced Dipole Interactions

Even molecules that have no net dipole moment, such as carbon dioxide or carbon tetrachloride, can have a transient dipole induced in them when they are brought into an electric field. Electronic interactions of this type are called *ion-induced dipole* interactions, and the potential of the interaction will depend on its *polarizability* α, or the ability of the neutral molecule to be induced into a dipole by an external electric field:

$$U_{i\text{-}id} = \left(\frac{q^2\alpha}{8\pi\varepsilon_o\varepsilon^2 r^4}\right) \tag{9.3}$$

9.3.4 van der Waals Interactions

When molecules are in very close proximity, attractive forces called *van der Waals* interactions occur. van der Waals forces are cohesive forces that vary with respect to distance as $1/r^6$. It is generally useful to subdivide the van der Waals interactions into three types, all derived from electrostatic considerations. These include permanent dipole–dipole, permanent dipole-induced dipole forces, and induced dipole-induced dipole interactions. van der Waals forces are considered to be the molecular explanation for the cohesive energies of liquids and are similar in magnitude to the enthalpies of vaporization of most liquids, approximately –41.84 kJ/mol. They can become dominant in reactions where close fitting and proximity are important. Each of these terms will be defined.

9.3.4.1 Dipole–Dipole Interactions

Dipole–dipole interactions occur when molecules with permanent dipoles interact. In dipole–dipole interactions, the orienting forces acting to align the dipoles will be countered by randomizing forces of thermal origin. These terms are reflected in Eq. (9.4). This equation indicates that dipole–dipole interactions are sensitive to both temperature and distance:

$$U_{d\text{-}d} = -\frac{2}{3}\left(\frac{\mu_1^2\mu_2^2}{(kT(4\pi\varepsilon_o)^2r^6)}\right) \tag{9.4}$$

9.3.4.2 Dipole-Induced Dipole Interactions

Permanent dipoles can induce a dipole moment in a neutral molecule in a fashion similar to that discussed above for ion-induced dipole interaction. This *dipole-induced dipole* interaction depends on the polarizability of the neutral molecule but is not sensitive to the thermal randomizing forces. This is reflected by the absent *kT* term in Eq. (9.5):

$$U_{d\text{-}id} = -\frac{2\mu_1^2\alpha^2}{16\pi^2\varepsilon_o r^6} \tag{9.5}$$

9.3.4.3 Induced Dipole-Induced Dipole or London Forces

The dipole moment is a time-average measurement. If a snapshot were taken of any neutral molecule at an instant of time, there would be a variation in the distribution of the electrons in the molecule. At this instant, the “neutral” molecule would have a dipole moment. This instantaneous dipole is capable of inducing an instantaneous dipole in another neutral molecule; thus are born *induced dipole-induced dipole* forces. These forces are also called *London* or *dispersion* forces. Dispersion forces fall off very rapidly with distance but can be quite significant for molecules in close proximity. The deformability of the electron clouds, as reflected by α, is obviously important in these interactions. The rigorous calculation for dispersion forces is quite involved, but an adequate approximation can be written as

$$U_{id\text{-}id} = -\frac{f(I)\alpha^2}{16\pi^2\varepsilon_o r^6} \tag{9.6}$$

The term *f*(*I*) is a function of the ionization energies of the two molecules and is equal to

$$f(I) = \frac{3I_1I_2}{2(I_1 + I_2)} \tag{9.7}$$

The overall interactional energy for van der Waals forces can be written in terms of these three electrostatically derived interactions:

$$U_{\mathrm{vdW}} = -\frac{2}{3}\left(\frac{\mu_1^2\mu_2^2}{kT\,(4\pi\varepsilon_{\mathrm{o}})^2\,r^6}\right) - \frac{2\mu_1^2\alpha^2}{16\pi^2\varepsilon_{\mathrm{o}}r^6} - \frac{f(I)\alpha^2}{16\pi^2\varepsilon_{\mathrm{o}}r^6} \tag{9.8}$$

The *attractive interactional energy* represented by the van der Waals forces is often written in the form:

$$U_{\mathrm{interaction}} = -\frac{A}{r^6} \tag{9.9}$$

where *A* is a different constant for each molecule. A graph of the van der Waals interactional energy would show that the attractive force increases rapidly as molecules get closer (Fig. 9.3) until the molecules actually contact one another. At that point the interactional energy must go instantly to infinity. In fact, before the molecules try to occupy the same space, another event occurs that leads to a repulsive force. This is a repulsion that occurs between the electron clouds of the molecules as they approach one another. The repulsion is due to the Pauli exclusion phenomenon and

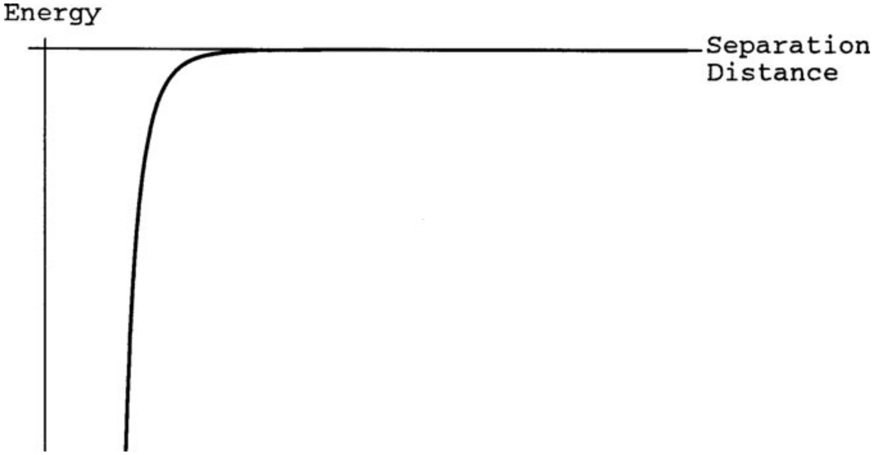


Fig. 9.3 The van der Waals interactional energy increases rapidly as molecules get closer until the molecules actually contact one another, at which point the interactional energy must go instantly to infinity

is not due wholly to the electrostatic repulsion between the electron clouds. This electron repulsion must be added into the formula for interactive energy, Eq. (9.9), and the interactive energy is then usually written in the form of the *Lennard–Jones potential*:

$$U_{\text{Lennard-Jones}} = -\frac{A}{r^6} = \frac{B}{r^{12}} \tag{9.10}$$

A graph of the Lennard–Jones potential is shown in Fig. 9.4.

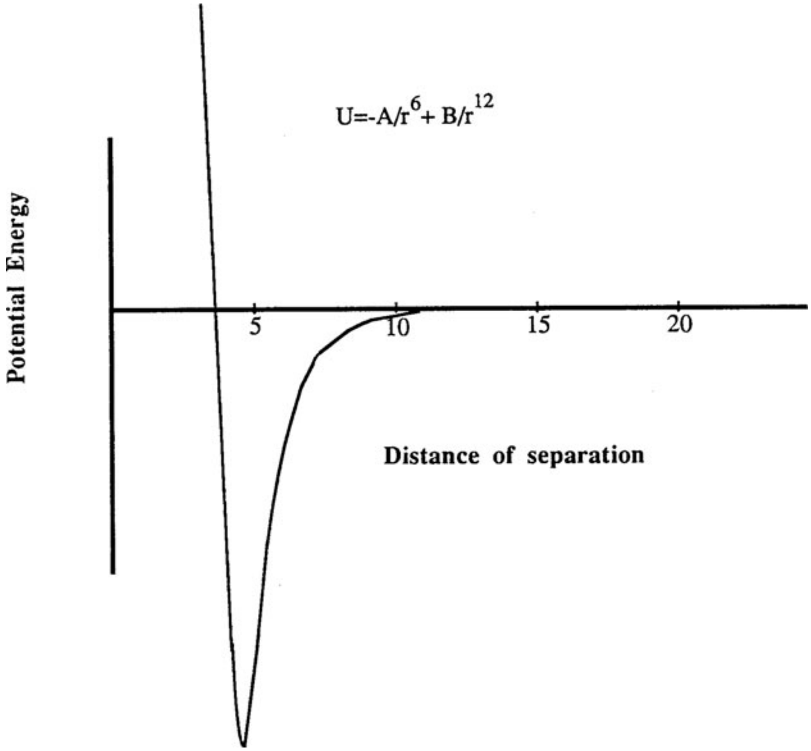


Fig. 9.4 The Lennard–Jones potential

9.4 Covalent Bonds Involve a True Sharing of Electrons Between Atoms

The covalent bond is a chemical bond that is formed when two atoms share electrons in contrast to the electrostatically dominated interactions in the ionic bonds. The modern description of the covalent bond is the subject of valence theory. Valence theory is important in describing the properties of chemical compounds and elements and explains the number, length, energy, and three-dimensional shape of the covalent bonds between atoms. There are a variety of abstract models that have been devised to describe the covalent bond. Like most abstractions each has its limitations as well as strengths. We will briefly review the models (with which most students are to varying degrees familiar) in preparation to their application to covalent bonds of biological interest.

9.4.1 Lewis Structures Are a Formal Shorthand that Describe Covalent Bonds

Prior to the full development of quantum theory, G.N. Lewis proposed that the covalent bond consisted of a pair of shared electrons, one from each atom in a molecule, and that each atom would share electrons until its outer shell was a completed octet with a configuration similar to the nearest noble gas. Although much of our present-day understanding of the covalent bond is based on quantum mechanical treatments, the Lewis structures remain a fundamental shorthand for many chemical problems. This is especially true in biochemical studies. Lewis structures do not provide geometrical or structural information but are rather shorthand for the arrangement of bonds. In brief the method for writing a Lewis structure is as follows:

1) The atoms are drawn, each having its periodic complement of electrons in the outer valence ring.
2) The atoms are arranged as in the molecule.
3) One electron pair (**:**) is added between each bonded atom.
4) The remaining electron pairs are arranged either as multiple bonds or as lone pairs.
5) Each bonding pair is replaced with a line. Lone pairs are left as (**:**).

Modern usage of Lewis structures includes the idea of resonance hybrids, which are multiple ways in which a structure can be rendered with the only difference being the location of multiple bonds. This description implicitly recognizes the concept of electron delocalization, which is a quantum mechanical result.

9.4.2 VSEPR Theory Predicts Molecular Structure

The Lewis theory of covalent bonding does not provide any significant tools for determining atomic or molecular structure. The valence shell electron pair repulsion model (VSEPR) focuses on a central atom and its complement of valence shell electron pairs. These electron pairs occupy one- or two-center orbitals that are *localized* and have relative size and shape. The arrangement of the orbitals around the central

Table 9.1 Arrangements of atoms in VSEPR theory

Electron pair number	Ideal geometry
2	Linear
3	Trigonal planar
4	Tetrahedral
5	Trigonal bipyramidal
6	Octahedral
7	Pentagonal bipyramidal

atom predicts the geometry of the overall molecule. The predicted arrangements are listed in Table 9.1. The rules for applying this model system are as follows:

1) The preferred arrangement is the one that maximizes the separation of the electron pairs.
2) A non-bonding pair of electrons (a lone pair) takes up more room than a bonding pair.
3) The size of the bonding electron pair decreases as the electronegativity of the ligand increases.
4) The electron pairs of a multiple bond take up more room than a single-bonded electron pair.

VSEPR theory predicts with complete accuracy that an AB_2 molecule is linear, an AB_3 molecule is trigonal planar, an AB_4 molecule is tetrahedral, and an AB_2E_2 molecule ($E =$ lone pair) will be bent.

9.4.3 Molecular Orbital Theory Is an Approximation to a Full Quantum Mechanical Treatment of Covalent Interactions

Ideally the theoretical treatment of molecular structure as defined by bond angles, lengths, and energies would be treated by ab initio quantum mechanical molecular orbital calculations that would be detailed and complete. The difficulty with these details, however, is that even the simplest molecule is made up of three particles and so the Schrödinger equation cannot be solved analytically but must be approximated. The fundamental approximation of all molecular orbital methods (and it is an approximation not an abstraction) is the *Born–Oppenheimer approximation*. This approximation takes into account the great differences in mass between the nuclei and the electron in a molecule. Thus the inertia of movement for a nuclear mass is very much longer than that of an electron. If the nuclear positions are moved, the electronic response is essentially instantaneous. Therefore, we ignore nuclear movement and instead treat the molecule as frozen in its nuclear tracks and solve for the kinetic and potential energy of the electrons. The frozen nuclear configuration is called a *molecular conformation*. The full solution requires that multiple molecular conformations are sampled, and a potential energy surface (or curve for a diatomic molecule) is constructed that shows the relationship between the conformation and the energy. The equilibrium conformation corresponds to the minimum of the energy surface. Figure 9.5 shows a potential energy curve for a diatomic molecule in which the energy is dependent on the internuclear distance.

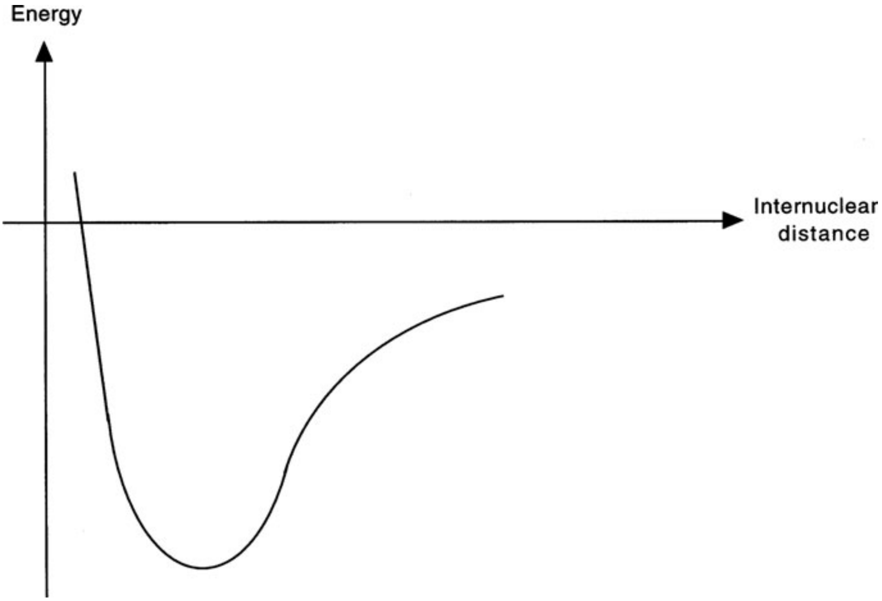


Fig. 9.5 A potential energy curve for a diatomic molecule in which the energy is dependent on the internuclear distance

Just as we saw in the case of the many-electron atom, the polyatomic molecule cannot be solved exactly even with the Born–Oppenheimer approximation. The only molecule that yields to analytical solution of the Schrödinger equation is the hydrogen ion molecule, H_2^+. In a fashion similar to the extrapolation from the atomic orbitals of the hydrogen atom to polyatomic orbitals, the solution of the H_2^+ molecule gives a set of molecular orbitals that have been used as the approximate basis for polyatomic biomolecules. Consider the picture of the two lowest energy molecular orbitals of H_2^+ as the two nuclei approach one another (Fig. 9.6). The lines around the nuclei are lines of electronic isodensity, in other words, they represent the probability of an electron being found in a particular region of space. The difference between the two orbitals lies in the occurrence of a node or a region of zero probability lying between the nuclei in the antibonding orbital compared to a region of higher electron localization in the bonding orbital. Thus in the bonding orbital the accumulation of electron probability between the nuclei leads to bonding in contrast to the exclusion of probability which represents antibonding (Fig. 9.7).

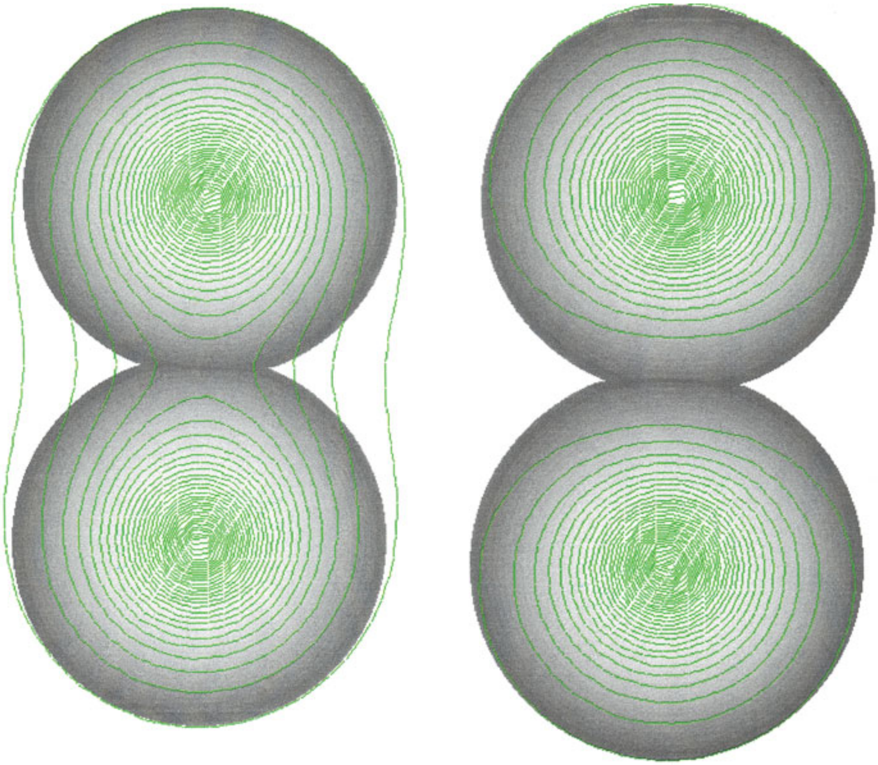


Fig. 9.6 Illustration of the two lowest energy molecular orbitals of H_2^+ . As discussed in the text, the nucleus on the *left* is *A* and on the *right* is *B*

Figure 9.8 illustrates the energies of these two orbitals. It is easy to see why one leads to bonding and the other does not.

When it comes to "exact" solutions we are now out of answers, which is not very practical for biophysical problems. The Hamiltonian for the hydrogen molecule, H_2 (see box), is analytically unsolvable by the Schrödinger equation. As we can appreciate by considering the interactions of two electrons and two nuclei, the problem becomes very complicated very quickly. Any molecule equally or more complex must be solved by one of two approximate methods. These two methods are the *molecular orbital*, also known as, the *linear combination of atomic orbitals molecular orbital* method (LCAO-MO) or the *valence bond* method. Each of these approximations derives from a different abstraction of the bonding nature of molecules. In the molecular orbital method the electrons are viewed as belonging to the entire molecule and therefore the individual bond is de-emphasized. The quantum mechanical wavefunction is written in terms of the molecule. In valence bond theory, the electron pair is viewed in terms of a localized bond, but the location of these bonding electrons is described by a wavefunction.

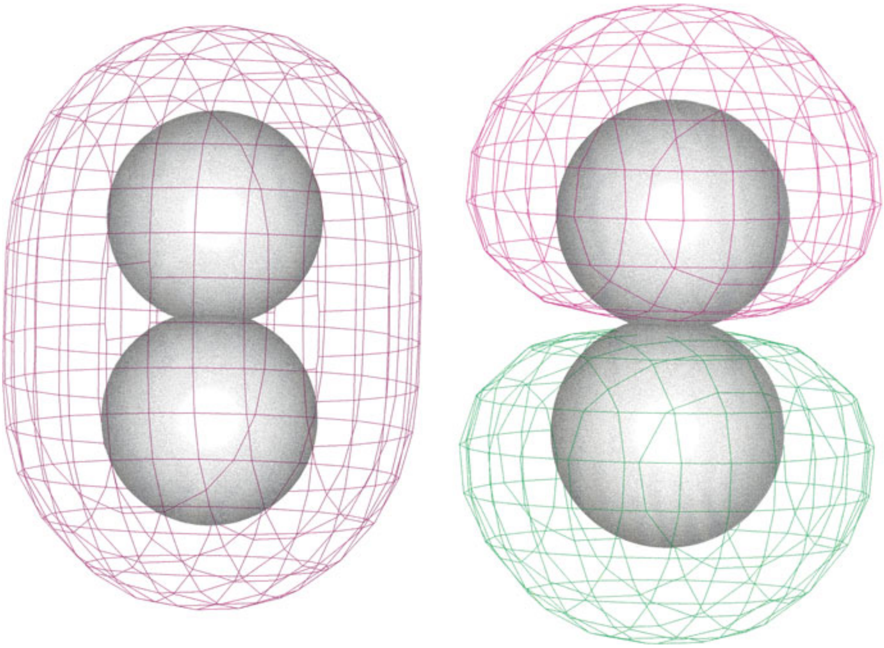


Fig. 9.7 Three-dimensional isosurfaces of the (*A*) bonding and (*B*) antibonding orbitals of H_2^+. The nodal plane in *B* can be clearly seen. There is no probability of an electron being found at this node, so bonding cannot occur in this orbital

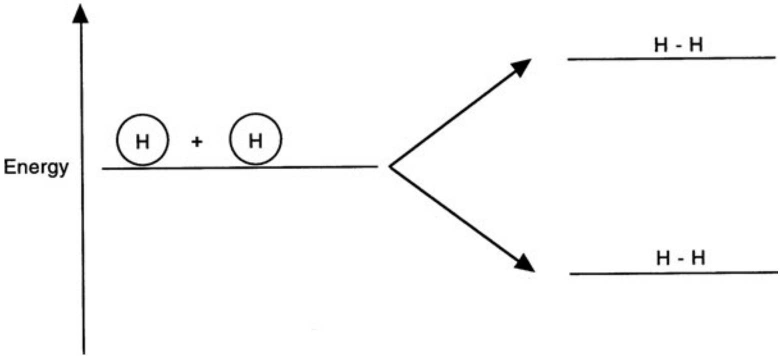


Fig. 9.8 Molecular orbital energy diagram for the H_2^+ atom

9.4.3.1 Linear Combination of Atomic Orbital's Molecular Orbital Method

Molecular orbital theory is most widely used in studies of small inorganic molecules and *d*-metal complexes as well as solid state systems. The vocabulary of the theory is widely dispersed throughout chemistry and biochemistry, though it probably finds the most direct application to organometallic compounds such as the metalloenzymes and the proteins involved in energy transduction and electron transfer. The LCAO-MO method builds a molecular orbital by adding together the atomic orbitals of the atoms from which the molecule is constructed. We have already seen this approach in our discussion to the "exact" solution of the H_2^+ molecule. Reexamine Fig. 9.7a. We see that near either nucleus the wavefunction of the electron closely resembles an atomic 1s orbital. In quantitative words we could say,

1) When the distance between the electron and nucleus A is small, the distance from nucleus B will be large and the wavefunction will be almost entirely $\psi_{1s}(A)$.
2) Alternatively, when the electron is near B, the function should reflect the opposite orbital, $\psi_{1s}(B)$.
3) The overall wavefunction for the molecular orbital should approach $\psi_{1s}(A)$ in proportional relation to the distance from A, and it should approach $\psi_{1s}(B)$ in proportional relationship to the distance from B.
4) $\psi_{1s}(A) \rightarrow 0$ as the distance r between the electron and the nucleus A increases and $\psi_{1s}(B) \rightarrow 0$ as the distance r between the electron and the nucleus B increases.
5) We can accomplish these simultaneous cases by combining the terms with an appropriate weighting factor (equal to each other) so that the molecular orbital is a weighted mixture of both atomic orbitals.

We can write these words more succinctly in terms of an equation:

$$\psi_{\mathrm{mo}} = \psi_{1s}(A) + \psi_{1s}(B) \tag{9.11}$$

This sum is the linear combination of the atomic orbitals. But the following expression also accurately describes these words:

$$\psi_{\mathrm{mo}} = \psi_{1s}\,(A) - \psi_{1s}\,(B) \tag{9.12}$$

Having two equations makes sense since we would expect to form one molecular orbital for each atomic orbital combined. From the standpoint of physical interpretation, we are interested in the localization of the electrons in the molecular orbitals just as we were in the atomic orbitals. If we apply the Born interpretation and square the wavefunction from Eq. (9.12), we can find the probability of finding the σ electron at any point in the molecule for the first of these two orbitals:

$$\begin{aligned}\psi^2 &= \{\psi_{1s}(A) + \psi_{1s}(B)\}^2 \\ &= \psi_{1s}(A)^2 + \psi_{1s}(B)^2 + 2\psi_{1s}(A)\psi_{1s}(B)\end{aligned} \tag{9.13}$$

Thus the probability of finding the electron near each nucleus is just as it is for an atomic orbital (as we expected). Furthermore, there is an additional component, which is the region of mixture. This is the region of overlap. If we think of this relationship in terms of waves, the overlap corresponds to a region of *constructive* *interference*. Now we square the wavefunction from Eq. (9.12) to find the electron distributions for the second orbital:

$$\begin{aligned}\psi^2 &= \{\psi_{1s}(A) - \psi_{1s}(B)\}^2 \\ &= \psi_{1s}(A)^2 + \psi_{1s}(B)^2 - 2\psi_{1s}(A)\psi_{1s}(B)\end{aligned} \tag{9.14}$$

Again the probability of finding the electron near each nuclei is just as it is for an atomic orbital (as we would expect). However, the region of mixture in this orbital corresponds to a region of *destructive interference*. There is a node of zero probability in the second orbital, a region the electron is unlikely to occupy. In the first orbital there is an accumulation of electron distribution between the nuclei, while in the second orbital there is exclusion. Thus the LCAO-MO method gives us a bonding and antibonding orbital, a result similar to the outcome of the Schrödinger equation. In our discussions we will not need to calculate molecular orbitals, but we will benefit from being able to draw diagrams of the molecular orbitals. We should also have some way of writing out the energy levels associated with each orbital.

If we now draw these two atomic orbitals and bring them together at a certain distance, it is clear that there is a region of overlap that does not have atomic orbital nature but rather a mixture of the two. The orbital formed by this procedure is called a σ *orbital*. The important quality of a σ orbital is that it is cylindrically symmetrical with respect to the internuclear axis. Any orbital that is cylindrically symmetrical is called a σ orbital. (The similarities to the symmetry of the atomic *s* orbital are clear and an electron that occupies a σ orbital has zero angular momentum around the internuclear axis.) We will draw these orbitals as being radially symmetrical and then combine them. In Eqs. (9.11) and (9.12) when the wavefunctions are both positive, we get constructive interference. When one wavefunction is positive and one is negative we get destructive interference and a node. We can use shading to distinguish the sign of the wavefunction. Both of these molecular orbitals are σ orbitals, the bonding orbital is of lower energy and named 1σ, while the antibonding orbital is named $2\sigma^*$ where the * represents the antibonding nature of the orbital. Figure 9.9 shows the geometry of these two molecular orbitals.

A molecular orbital energy diagram can be drawn that relates the contributing atomic orbitals and the formed molecular orbitals. In these diagrams the energy

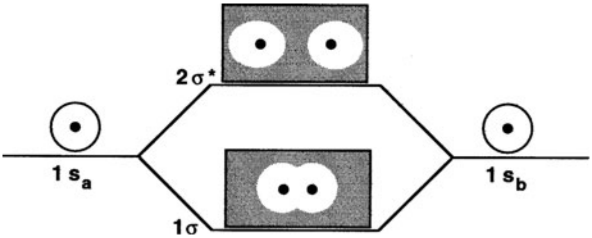


Fig. 9.9 The geometry of the bonding 1σ and the antibonding orbital 2σ* orbital

of the bonding and antibonding orbitals are drawn. The antibonding orbit often raises the energy more than a bonding orbital lowers the energy, and so an asymmetry is often found in the energy level diagram reflecting this energy difference. The separation of the energy levels corresponds to the equilibrium bond lengths. This relationship helps provide a physical equivalent to the idea of accumulation or exclusion. When electrons are of high enough energy to occupy antibonding orbitals, the internuclear distances are too far apart for bonding to occur and there is covalent bond breaking. In the case of simple molecules this will lead to dissociation, but in the case of complicated polyatomic molecules the breaking of certain bonds will change the shape, orientation, and reactivity of the molecule. The molecular orbital energy diagram gives a convenient method of discussing such changes.

Let us consider the formation of the molecular orbitals formed from atoms of period 2 which include carbon, nitrogen, and oxygen. A much larger number of atomic orbitals are now available for the LCAO-MO treatment. The available orbitals include the *core orbitals* which are inside of the valence orbitals (the closed 1*s* orbitals for period 2), the orbitals of the valence shell, the *valence orbitals*, and high-energy ground state unoccupied orbitals called *virtual orbitals*. As a first approximation the contributions of the core and virtual orbitals are not considered, though in certain cases this simplification must be reconsidered. We will not consider these contributions in our treatment. The valence orbitals in the second period are 2*s* and 2*p* orbitals. The *p* orbitals are oriented along the *x*-, *y*-, and *z*-axis and by convention the *z*-axis is the internuclear axis. We will expect for a diatomic molecule of this period eight molecular orbitals. We can see quickly that the 2*s* and the $2p_z$ orbitals are oriented with cylindrical symmetry along the internuclear axis and, thus, will form σ bonds. We would expect these four σ forming orbitals to admix together in some proportion so that each of the σ bonds has some *s* and *p* character, and in fact this is what happens. However, there is a greater *s* character to the first two σ bonds, 1*s* and $2s^*$, and greater *p* character to 3σ and $4\sigma^*$. The remaining *p* orbitals, p_x and p_y, are perpendicular to the internuclear axis and are not cylindrically symmetric with respect to the *z*-axis. Instead these orbitals overlap either constructively or destructively from the side. This interaction leads to π orbitals, a doubly degenerate bonding and a doubly degenerate antibonding pair. Two electrons in a π orbital form a π bond. The schematic for period 2σ and π bonds is shown in Fig. 9.10.

An important restriction in building these models lies in combining orbitals of equivalent symmetry with respect to the internuclear axis. Thus we can combine *s* and p_z orbitals and p_x and p_y orbitals, but we cannot cross-combine them. This is required because the increase in bonding accumulation must be offset by an equal amount of antibonding exclusion. (Consider the implications to the universe if this restriction were not obeyed.) This condition allows consideration of the role of *d* orbitals in molecular orbitals of the period 2 elements. We will discuss the relationship of symmetry and molecular orbitals again when we consider molecular spectroscopy.

The molecular orbital energy level diagrams provide a great deal of information that is important in understanding the properties of biological molecules. For

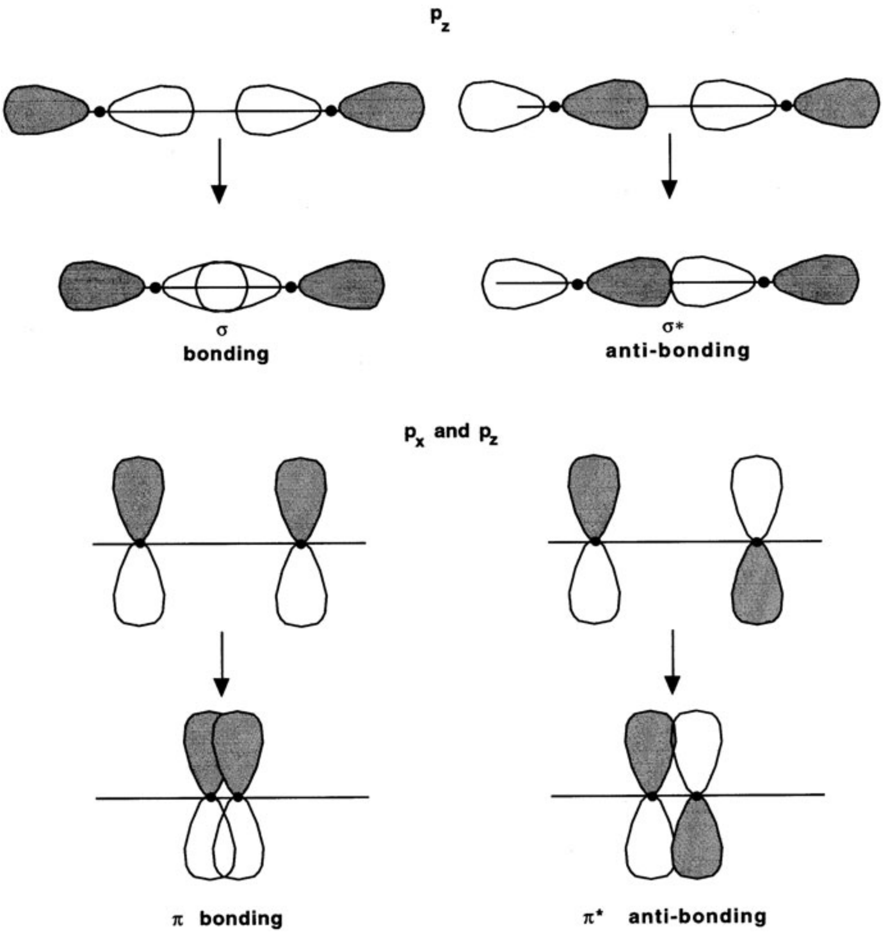


Fig. 9.10 The schematic for period 2σ and πbonds

example, the bond strength in a molecule can be understood in terms of these diagrams. Bond properties tend to be quite independent of the molecular and environmental conditions in which a bond is found. Consequently, understanding the nature of bonding makes available a very efficient abstraction that provides a great deal of information about molecular structure at least at the small molecule level. The bond strength represents the combination of the occupied bonding and antibonding orbitals. This combination is called *bond order*. In the case of biologically important diatomic molecules such as O_2 and N_2, the bond order is determined by the formula $b = \frac{1}{2}(n - n^*)$, where n is the number of electrons in each orbital. The bond order, bond length, and dissociation energy for the biologically important diatomic molecules are shown in Table 9.2. This analysis shows that if the

Table 9.2 Properties of diatomic species

Diatomic species	Bond order	Bond length Å	Dissociation energy (kJ/mol)
He	0	–	–
H_2	1	0.74	435.1
Li_2	1	2.673	104.6
O_2	2	1.207	497.9
N_2	3	1.094	945.6
F_2	1	1.42	154.8

antibonding orbitals are equally occupied, covalent bond formation does not occur. Furthermore, bond strength is proportional to bond order and inversely proportional to bond length.

The molecular orbitals for heteroatomic molecules are considered in the same basic fashion. In principle, the only major modification is to consider the inequality of the coefficients of the wave equation that represents the linear combination:

$$\psi_{\text{mo}} = c_{\text{A}}\psi(A) + c_{\text{B}}\psi(B) \tag{9.15}$$

In keeping with the Born interpretation, the proportion of ψ (A) in ψ_{mo} will be $(\text{c}_A)^2$ and the contribution of $\psi(B)$ can be treated similarly. For a nonpolar covalent bond $c_A = c_B$. For an ionic bond $c_{\text{B}} = 0$ and $c_{\text{A}} = 1$. The atomic orbital with lower energy, which belongs to the more electronegative atom, will make the larger contribution to the bonding orbital.

9.4.3.2 More on the Molecular Bond

What is the essence of the molecular bond in terms of the energy concepts that contribute to it?

A conventional perspective treats the problem in terms of energy. This argument proceeds as follows. The energy rise seen at short internuclear distances arises because of repulsion of the nuclei keeping them from collapsing into a single nuclear entity (an He^+ isotope). The bonding energy may be viewed as the favorable energy that arises from the presence of the electron between the two nuclei that acts to reduce the energy due to repulsion. The antibonding curve shows the unfavorable electrostatic effect of excluding the electron from the internuclear region.

An alternative perspective is to consider the problem in terms of forces being exerted by the electron on the nuclei. In this case we recognize that the electron exerts an electrostatic force on each nucleus proportional to $\dfrac{e^2}{r_n^2}$. This force can be resolved into forces perpendicular and parallel to the internuclear axis. When the electron is outside a nucleus, it acts to separate the nuclei from one another. When it is between the nuclei, it draws the nuclei closer together. This analysis can be applied to the space around the nuclei and a boundary surface can be drawn inside which electrons will hold the nuclei together.

Both of these arguments have a very classical flavor, and it should come as no surprise that other more quantum-like explanations can be made. It can be shown, at least for the H_2^+ molecule, that when the bonding orbital forms, the wavefunction is distorted by shrinking and fitting more closely to the nuclei, while the antibonding orbital leads to an expansion of the wavefunction near the nuclei. The changes in the wavefunction make it more likely that the electrons are found between the nuclei than outside them. It is not the electrons that glue the molecule together then but rather the compressed wavefunction which allows shrinkage of the orbitals.

9.4.3.3 Valence Bond Theory

Like molecular orbital theory, valence bond (VB) theory is an attempt to describe molecular bonding, structure, and energy based on electronic distribution. Both use the quantum mechanical wavefunction to find the distribution; however, VB theory starts with the electron pairs making up the chemical bond as a fundamental assumption. In VB theory, the pairs of electrons are established and then a wavefunction is sought that allows the pair of electrons to be found on either atom. All of the possible arrangements of electrons making up a bond are assembled into an ensemble, and the overall picture of the bond is dominated by the most probable electronic structure. This method is in sharp contrast to the MO approach, in which we have seen the orbital picture constructed from atomic orbitals followed by the step-by-step insertion of electrons into these orbitals. Thus the VB method focuses on the chemically intuitive concept of the bond being built from electron pairs, while MO theory views the electrons as being completely delocalized. An important difference that results from these divergent starting points is that MO theory does not naturally account for *electron correlation*, or the tendency of electrons to keep apart from one another. Thus electron correlation is under emphasized in MO theory. VB theory, on the other hand, insists on electron pairing and, thus, overemphasizes the role of electron correlation in the distribution picture. MO theory has received greater attention over the years mainly because it is easier to solve problems from a computational standpoint. But the ideas arising from VB theory are applied widely in bio-organic chemistry and are useful abstractions for understanding the structure and properties of biological systems.

Bonding in VB theory occurs in cylindrically symmetric σ bonds and π bonds just as in MO theory. Because all bonds must occur with a pair of electrons, the idea of promoting electrons to higher atomic orbitals so that molecular bonding can occur must be introduced in VB theory. The tetravalent structure of carbon does not seem reasonable given its ground state configuration of $1s^2$, $2s^2$, $2p_x^1$, $2p_y^1$, which implies a maximum of two bonds. But if one of the $2s$ electrons is promoted to the $2p_z$ orbital, then there are four electrons available for bonding. The energy needed for the *promotion* of the electron is recovered by the energetically favorable configuration of the molecular structure. This energy includes the reduction in electron repulsion energy that comes about when two electrons occupy the 2*s* orbital, as well as the two new bonds that can be formed.

The orbitals that are available for bonding in VB theory are a mixture of their component atomic orbitals. These orbitals are called *hybrid* orbitals. In the case of carbon there are four hybrid bonding orbitals, each comprised of a combination of the *s* and *p* atomic orbitals. We call hybrid orbitals composed of three *p* and one *s* orbitals sp^3 hybridized orbitals. These orbitals are equivalent but are oriented differently in space such that the four orbitals point to the corners of a regular tetrahedron. The shape of the orbital is shown in Fig. 9.11. The structure of the nonpolar hydrocarbon methane is easily understood in terms of VB theory. It is described as an sp^3 hybridized carbon atom that forms equivalent bonds with a hydrogen atom at each point of the tetrahedron. If another sp^3 hybridized carbon bonds at one of the points of the tetrahedron, a bond will be formed between them that is cylindrically symmetrical and the carbon atoms will be free to rotate around the bond without hindrance (assuming everything else is equal, which is often not the case). VB theory further demonstrates its usefulness as an abstraction when considering the formation and properties of double bonds. According to VB theory, a double bond formed between two carbon atoms requires different hybridization. In this hybrid orbital one *s* and two of the three *p* atomic orbitals (p_z and *py*) are combined. The constructive and destructive interference of the wavefunctions gives rise to three hybrid orbitals designated sp^2. These orbitals lie in a single plane, each at a mutual angle of 120°. The remaining p_x orbital is not included and it lies perpendicular to the plane of the sp^2 orbitals. If two sp^2 orbitals from two carbon atoms now bond, they will share electrons from the sp^2 orbital and so will lie in a plane with a bond angle

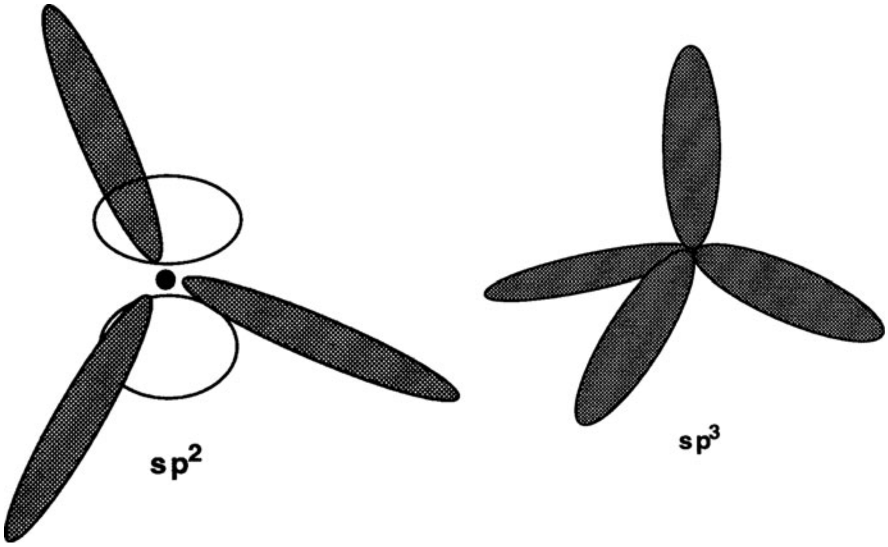


Fig. 9.11 The hybrid orbital structures: (**a**) The sp^2 hybrid orbital. (**b**) The sp^3 hybrid orbital.

of 120°. If the only interaction were the formation of the cylindrically symmetric σ bond, these two planar carbon atoms could freely rotate around one another. However, when all of the atoms are coplanar, the orthogonal p_x orbitals are able to interact and share their valence electrons. A second bond is formed, the π bond. The required two pairs of bonding electrons to form a double bond make the bond rigid (because it can only exist in limited angular configurations), make the double-bonded molecules have a local structure of coplanar 120° bond angles, and make the internuclear distance shorter than the bond between two sp^3 orbitals. This description is essentially a description of the molecule ethylene, $H_2C{=}CH_2$ (Table 9.3). Different types of hybrid orbitals can be formed depending on the combination of orbitals. The geometries of hybrid orbitals are described in Table 9.4.

Table 9.3 Comparison among carbon–carbon bonds of ethane, ethylene, acetylene

	Bond angle (degrees) <CCH	Bond length (Å)	Energy (kJ/mol)
Ethane	111.17°	1.54 Å	347.3
Ethylene	121.3°	1.34 Å	610.9
Acetylene	180°	1.2 Å	896.8

Table 9.4 Hybrid orbital geometries

Hybridization pattern	Coordination number	Shape
sp	2	Linear
sp^2	3	Trigonal planar
sp^3	4	Tetrahedral
sp^3d	5	Bipyramidal
sp^3d^2	6	Octahedral

9.5 Hydrogen Bonds Are a Unique Hybrid of Interactions and Play a Fundamental Role in the Behavior of Biological Systems

The *hydrogen bond* is of paramount importance in aqueous systems. The bond is likely composed of electrostatic, covalent, and resonance components. However, the electrostatic interactions are probably dominant. Because of its mixed nature, we place it at the end of our discussion of bonding.

Hydrogen bonds consist of an interaction between the covalent *X*–H bond of one molecule which is electronegative and the lone electron pair of another atom, *Y*, also electronegative. Generally *X* and *Y* are considered to be O, N, or F. Most early research on the hydrogen bond was performed in condensed phases, with infrared and Raman spectroscopy providing information of the bonding patterns involved.

NMR experiments provided dynamic information on formation and dissociation equilibria and rates. The geometry of the H-bond shows a linear gluing of two heavier atoms by a hydrogen atom. The quantum mechanics of bond formation requires a pair of electrons available on the *Y* atom for interaction. Thus, lone pairs are considered to point toward the bridging hydrogen. It is commonly held that the bridging distance should be smaller than the sum of the van der Waals radii. The energetics of the H-bond has the bond strength between covalent and van der Waals interactions, usually between 8 and 60 kJ/mol. When one of the two subunits is electrically charged, the interactional energy can be as high as 165 kJ/mol. Spectroscopic evaluation of the *X*–H stretching is in the infrared band and shifts to the red and broadens with formation of the H-bond.

In the last several decades quantum mechanical calculations have allowed a more considered evaluation of the roles played by electrostatic attractions and steric repulsions of the electron clouds. The classic hydrogen bond such as the one found in water, which is linear with an interoxygen distance of 0.30 nm, is clearly well defined. However, there are other cases such as $H_3N \cdots {}^+HNH_3$, which give rise to "strong" H-bonds in which ionic character begins to swamp the quantum effects responsible for the H-bonding. To amplify this point, consider the case of $ClH \cdots NH_3$ which is linear and is a classic hydrogen bond. However, if the proton shifts so that the result is $Cl^- \cdots {}^+HNH_3$, the ion–ion interaction clearly swamps the hydrogen bonding character even though the relationship remains linear. Similar arguments at the weak end of the continuum, where a nonpolar methane hydrogen may be pointing toward the lone pair of an oxygen, make the idea of the hydrogen bond somewhat less distinct.

The biological hydrogen bond is a relatively weak bond of between –20 and –30 kJ/mol. The strength of the bond increases with increasing electronegativity and decreasing size of the participating atoms. Thus, hydrogen bonds can exist in a number of compounds, not just in H_2O. The study of hydrides, amides, hydrogen halides, hydrogen cyanide, and ammonia has helped to clarify the nature of the hydrogen bonds which link compounds of these types into linear polymers. However, our focus will be the hydrogen bonds in water, in aqueous solutions, and within the solutes in such solutions. The important hydrogen bonds in biological systems are summarized in Table 9.5.

Table 9.5 Characteristics of H-bonds of biological importance

X–H ··· Y	Bond distance (Å)	Bond energy (kJ/mol)
O–H ··· O	0.270	–22
O–H ··· O–	0.263	–15
O–H ··· N	0.288	–(15–20)
N^+–H ··· O	0.293	–(25–30)
N–H ··· O	0.304	–(15–25)
N–H ··· N	0.310	–17
HS–H ··· SH_2		–7

9.6 Biological Systems Are Made from a Limited Number of Elements

What are the biological materials to our study of physical theories will be applied? Biological chemistry is the chemistry of biomolecules both in solution and at surfaces. The molecules of concern can be broadly arranged into a number of abstract categories based on their size, composition, and charge. This is often the approach used in considering the general physical interactions. We will need to define which molecules that we will consider. We summarize some aspects of their basic biochemistry here.

Small molecules of concern in biological systems include molecular water, ions, and small organic molecules including a wide variety of lipids, steroids, carbohydrates, amino acids, peptides, and nucleic acids. Larger molecules or macromolecules are constructed of polymerized chains or associated arrays of these smaller molecules. Amino acids polymerize to form proteins. Nucleic acids polymerize to form RNA and DNA. Carbohydrates polymerize to form complex sugars and starches. Lipids associate in the presence of an aqueous phase into membranes. Bulk water is, to a reasonable approximation, a hydrogen-bonded polymer of H_2O molecules. The rich diversity of biological organisms results from the fact that the relatively small group of small molecules can combine in a seemingly infinite variety of distinct forms as larger molecules. The elemental components making up biological systems are summarized in Table 9.6. Table 9.7 then summarizes the role of self-association and polymerization in each of the classes of biomolecules, demonstrating how new families of function can be derived by supramolecular organization.

Table 9.6 Summary of the constituent elements in biological systems

		Relative numbers	
	Ground-state structure	(per 1000 atoms)	
	1 2 3 4 5	(trace elements	
Element	*s s p s p d s p d f s p d*	given in mol/l)	Examples of role
H	1	605.70	H_2O, p, n, c, l
C	2 2 2	106.67	p, n, c, l
N	2 2 3	24.36	p, n, c, l
O	2 2 4	256.77	H_2O, p, n, c, l
F	2 2 5	0.74	Matrix of bone
Na	2 2 6 1	0.11	Electrolyte
Mg	2 2 6 2	0.009	Chlorophyll, ATPase
Si	2 2 6 2 2	1.34	Mostly in plants
P	2 2 6 2 3	1.32	ATP, l, n
S	2 2 6 2 4	0.33	p, iron–sulfur enz
Cl	2 2 6 2 5	0.37	Electrolyte
K	2 2 6 2 6 1	22.6	Electrolyte
Ca	2 2 6 2 6 2	4 nM	Electrolyte, bones
Cr	2 2 6 2 6 5 1	1 nM	Glucose metabolism
Mn	2 2 6 2 6 5 2	18 μM	Hydrolases, oxidases

Table 9.6 (continued)

Element	Ground-state structure 1 2 3 4 5 *s s p s p d s p d f s p d*	Relative numbers (per 1000 atoms) (trace elements given in mol/l)	Examples of role
Fe	2 2 6 2 6 6 2	0.1 nM	Heme, iron–sulfur enz
Co	2 2 6 2 6 7 2	20 nM	Transferases
Ni	2 2 6 2 6 8 2	16 μM	Urease
Cu	2 2 6 2 6 10 1	15 μM	Superoxide dismutase
Zn	2 2 6 2 6 10 2	1.6 μM	RNA polymerase
Se	2 2 6 2 6 10 2 4	7 nM	Glutathione peroxidase
Mo	2 2 6 2 6 10 2 6 5 1	40 nM	Xanthine oxidase
I	2 2 6 2 6 10 2 6 10 2 5		Thyroxine

p = peptides; n = nucleic acids; c = carbohydrates; l = lipids; enz = enzymes

Table 9.7 Supramolecular organization can generate new families of function

Class	Family	Function	Examples
Amino acids and proteins			
	Small peptides and amino acids		
		Cell signaling and control	Enkephalins, dopamine, adrenaline, thyroxine, glutathione
		Energy sources	
		Gluconeogenesis	Alanine, arginine, cysteine, glycine, serine, histidine
		Ketogenesis	Leucine, lysine, tryptophan
		Cellular defenses	Interleukins, lymphokines
	Proteins		
		Structural	Collagen, elastin
		Catalysts and enzymes	Catalase, elastase, hexokinase
		Signaling	Insulin, parathormone, growth hormone
		Movement	Actin, myosin, dynein, dystrophin
Carbohydrates			
	Mono and diglycerides		

Table 9.7 (continued)

Class	Family	Function	Examples
		Energy/carbon source	Glucose, fructose, sucrose, maltose
	Polymerization and polydisperse chains		
		Energy and carbon source	Glycogen, amylase
		Structural elements	Cellulose, glycosaminoglycans, mucopolysaccharides
		Cell recognition	N-acetyl neuraminic acid, galactose
Lipids			
	Monolipids		
		Intracellular signaling	Steroid hormones, vitamin D, retinoic acid
		Energy and carbon source	Triglycerides
	Micelles		
		Energy and carbon source	Chylomicrons
	Membranes		
		Intracellular signaling	Inositol-phospholipid signaling pathway
		Structural elements	Bilayer phospholipids
Nucleic acids			
	Components of high-energy currencies		ATP, GTP
	Components of signaling systems		cAMP, GTP
	Components of gene expression		Double-helical DNA, circular DNA, messenger RNA, transfer RNA, ribosomes

Further Reading

History of Chemistry

Partington J.R. (1989) *A Short History of Chemistry*. Dover Publications, New York.

General

For those who wish to review several important introductory topics without resorting to a textbook, and also for a different look at these traditional topics, the following series of articles on "Demystifying Introductory Chemistry" is recommended:

Gillespie R.J., Spencer J.N., and Moog R.S. (1996) Electron configurations from experiment, *J. Chem. Educ.*, **73**:617– 622.

Gillespie R.J., Spencer J.N., and Moog R.S. (1996) Bonding and molecular geometry without orbitals – the electron domain model, *J. Chem. Educ.*, **73**:622–627.

Gillespie R.J., Spencer J.N., and Moog R.S. (1996) An approach to reaction thermodynamics through enthalpies, entropies and free energies of atomization, *J. Chem. Educ.*, **73**:631–637.

Spencer J.N., Moog R.S., and Gillespie R.J. (1996) Ionization energies, electronegativity, polar bonds and partial charges, *J. Chem. Educ.*, **73**:627–631.

Intermolecular Forces

Campanario J.M., Bronchalo E., and Hidalgo M.A. (1994) An effective approach for teaching intermolecular interactions, *J. Chem. Educ.*, **71:**761–766. (Discusses the use of potential energy maps in understanding chemical properties and interactions. Examples are predominantly biochemical applications.)

Israelachvili J. (1992) *Intermolecular and Surface Forces*, 2nd edition. Academic Press, London. (The first section of this book provides a good exploration of the intermolecular forces in chemical systems.)

Bonding

Atkins P.W. and Friedman R.S. (1996) *Molecular Quantum Mechanics*, 3rd edition. Oxford University Press, New York.

Gillespie R.J. (1992) Multiple bonds and the VSEPR model, *J. Chem. Educ.*, **69:**116–120.

Pauling L. (1960) *The Nature of the Chemical Bond and the Structure of Molecules and Crystals: An Introduction to Modern Structural Chemistry*, 3rd edition. Cornell University Press, Ithaca, NY.

Scheiner S. (1994) Ab initio studies of hydrogen bonds: the Water Dimer paradigm, *Annu. Rev. Phys. Chem.*, **45**:23–56.

Problem Sets

1. Calculate the Coulomb interaction for Na^+ and Cl^- when separated by the distance of their atomic radii.
2. Calculate the magnitude of the thermal energy kT at 300 K and at 310 K.
3. Compare the energy of the Coulomb attraction from question two with the thermal energy at room temperature (300 K).

4. What separation is needed to make the Coulomb attraction energy approach the scale of the thermal energy at room temperature?
5. What is the energy $U(r)$ and force $F(r)$ predicted by the Lennard–Jones potential when two atoms are at their equilibirum (minimum) separation, $\mathrm{A} = 10^{-77}\ \mathrm{J\ m}^6$; $\mathrm{B} = 10^{-134}\ \mathrm{J\ m}^{12}$.
6. Mammals have very tightly controlled temperature ranges compatible with life. Temperatures below 27°C and above 41°C generally result in death probably from loss of activity of enyzme activity and normal receptor-agonist binding. Assume that the major contribution to these binding interactions is dispersion forces and hydrogen bonds. Further assume that the conformation of the enzymes and receptors does not change significantly at these temperatures. Calculate if the variation in thermal energy is likely to change the energy of interaction significantly.
7. Use a systems organizer (Chapter 2) to present the modeling systems of the Lewis structure, VSPER theory, and LCAO-MO theories – What are the emergent properties of each of these systems of abstraction? Compare the systems and discuss when one or the other is a model with greater goodness than another.
8. Comment on the implications to the universe if the following restriction in LCAO-MO theory were not found: An important restriction in building LCAO-MO models lies in combining orbitals of equivalent symmetry with respect to the internuclear axis. Thus, we can combine s and p_z orbitals and p_x and p_y orbitals, but we cannot cross-combine them.

Chapter 10
Measuring the Energy of a System: Energetics and the First Law of Thermodynamics

Contents

10.1 Historically Heat Was Thought to Be a Fluid or "the Caloric"

Thermodynamics is the study of heat and its transformation into mechanical work and vice versa. It is codified into three laws, each of which will be discussed in the following chapters. The first of these laws is axiomatic and is generalized from centuries of experience with observing energy transformations. The second law was derived theoretically: it considered what might be the maximum work available during the transfer of heat to a suitable engine. The third law is also derived and defines an absolute minimum for the entropy for certain systems.

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_10, © Springer Science+Business Media, LLC 2010

Countless experiences have shown that a motor may produce mechanical energy only if other energy is provided to it. The mechanical energy produced plus the frictional losses and heat lost in the transformation is always exactly equal to the energy put into the motor. The first law codifies this experience. It states that within the Universe, the conversion of energy is unlimited but the conservation of that energy is absolute:

$$E_{\text{Universe}} = E_{\text{system}} + E_{\text{surroundings}} \tag{10.1}$$

No restrictions of any sort are placed on the conversion process; the first law simply states that the total energy before and after a conversion is the same:

$$\Delta E_{\text{Universe}} = \Delta E_{\text{system}} + \Delta E_{\text{surroundings}} = 0 \tag{10.2}$$

$$\Delta E_{\text{system}} = -\Delta E_{\text{surroundings}} \tag{10.3}$$

For example, mechanical energy can be completely converted into heat if a motor turns a wheel in a body of water. J.P. Joule demonstrated this experimentally in 1845. It is true that all forms of energy can be completely converted into heat, and this conversion can be measured as a rise in the temperature of some material. Such a conversion is entirely consistent with the first law of thermodynamics alone.

Historically heat was thought to represent a unique physical phenomenon akin to electric charge. Heat was regarded as a fluid, the *caloric*, which penetrated matter and would be expected to flow from one body to another according to a set of rules. Classical thermodynamics was built on the expectation that the development of a physical model of heat flow should be approached in much the same manner as the successful studies of electricity. Careful observations of the natural behaviors of systems that involved heat and work were to be made. Experimental results would be obtained and empirically validated axiomatic principles or laws of thermodynamics would be established. This approach to discovering valid principles was practically successful, and today classical thermodynamics remains valid and important, useful for the practical applicability of the laws and elegant because it is in itself a formal system of significant grace and beauty. But because thermodynamics is a macroscopic modeling approach to the world (versus Chapter 2), the results of classical thermodynamics do not require any mechanistic understanding or specificity. This makes it at once both a powerful generalizable tool and somewhat arcane as we learn about it since almost everyone learns more easily when a mechanistic explanation is given because it is easier to visualize a more concrete (mechanical) example.

Was the search for the “caloric” principle successful? No. The search for a mechanistic “fluid of heat” comparable to the “fluid of charge” was futile since it turned out that heat is simply the disordered motion of ordinary matter and obeys the laws of mechanics (most generally, quantum mechanics). Thus the classical empirical laws of thermodynamics are related to mechanics in terms of the theoretical modeling and in principle can be derived from the laws of mechanics. This historical lesson is an important salve to the molecular yearnings of the modern biophysical chemist. Even in the twenty-first century, where the desire to achieve a microscopic understanding of biological systems is central, the value of thermodynamics is impressive. It is important exactly because it can be applied to systems for which we may have extremely limited or even incorrect molecular knowledge, a condition that applies to almost all systems of biological interest.

Thermodynamics is a central tool used in biophysical studies to evaluate the energy of a system of interest. Because it is concerned with macroscopic properties, a thermodynamic approach is particularly valuable in characterizing work and energy exchange in complex systems (e.g., cells or organisms) and also for determining the balance or position of substances involved in chemical equilibrium simply by knowing the properties of the substances. These properties are macroscopic observables (such as pressure, volume, number of molecules) and are linked in a deterministic fashion by an equation of state. This top-down approach placing the emphasis on macroscopic properties allows description of a subsystem that we postulate to be a good model of the entire natural system. The choice of observables and the linking equations of state represent the method of abstraction in thermodynamics. The abstract description of the thermodynamic world is at the level of macroscopic observables.

In thermodynamic study, a connection between the macroscopic/ phenomenological view and the microscopic/molecular view can be made through statistical mechanics and results in statistical thermodynamics. Therefore two approaches to thermodynamics exist, classical and statistical thermodynamics.

- In the classical or macroscopic formulation, a limited set of observables is found to be empirically accurate in the description of a state of the natural system. These are invariably found by measurement. Though the nature of our Universe is in fact stochastic, the classical properties take this inherent statistical behavior into account and classical approach would be accurate regardless of whether there is a stochastic or deterministic physical nature to our world.
- In a statistical thermodynamic formulation, the mechanical interactions are at the root of the linkages. A purely mechanical representation of a natural system usually entails so many degrees of freedom that it is impossible to keep track of all of them. We must resort to statistical methods to keep track of the interactions. The combination of a statistical treatment of the population of individual actions comprising the system leads to a determination of the macroscopic properties that were otherwise empirically measured in the classical approach.

10.2 The Thermodynamic Modeling Space Is a Systemic Approach to Describing the World

Our earlier treatment of a general system as being composed from elements, rules of relationship, and having a context space into which it is embedded with the result of emergent properties is given a special form in the thermodynamicist’s toolbox. The components necessary to formulate a thermodynamic treatment of an aspect of the natural world include the definition of the system, surroundings, boundaries, system properties, system state, and an equation of state. This is specialized language and its usage must be recognized when discussing thermodynamic formulations to prevent confusion.

10.2.1 Systems, Surroundings, and Boundaries

A thermodynamic *system* is a part of the physical universe of interest to an observer that simplified when it is placed under investigation. Fine details are smeared away and the system is viewed as completely defined by a small number of uniform properties (e.g., pressure, temperature, internal energy). The system is confined to a definite place in space by the *boundary* that separates it (the included part) from the rest of the Universe (the excluded part) or the *surroundings*. This boundary itself has specific properties that govern the relationship between the system and the surroundings. Each of these three is important and has equivalent importance to the proper treatment of the model.

Boundaries are classified according to their permeability to matter, heat, or work. All interactions between the system and its surroundings occur through the boundary. If the boundary prevents any exchange of matter, heat, or work between the system and the surroundings, an *isolated system* is said to be present. Because it can have no effect on the surroundings, an isolated system produces no observable disturbance as viewed from the surroundings. This condition is reminiscent of the question about the sound of a tree falling in a forest when no one is present to hear it. If the experimenter is in the surroundings, any changes that might occur in an isolated system would not be measurable. A system that is not isolated may be either *open* or *closed*. A system is *open* when material or matter, heat, and work pass across the boundary. A boundary that allows heat and work but not material to pass is termed *closed*. If a boundary will allow only work but no heat or matter to pass, the system is *adiabatic*.

10.2.2 Properties of a Thermodynamic System

The well-defined system with its boundaries and surroundings is described in terms of measurable quantities, that is, macroscopic observables. The measurable attributes are called the *properties of a system*. These properties are those physical attributes that are perceived by the senses or are made perceptible by instrumentation or experimental methods of investigation. Typical examples of thermodynamic properties are temperature, pressure, concentration of chemical species, and volume. Useful properties would have one or both of the following characteristics:

(1) Repeated measurements of the same property at equilibrium yield the same value. This is important because in kinetic studies it is the change in the value of a property as a system approaches equilibrium that is studied.

(2) Different methods for measuring the same property yield the same result. This ensures that the system is being described, rather than an artifact of the experimental procedures of measurement.

10.2.3 Extensive and Intensive Variables

Sometimes it is useful to differentiate between fundamental properties and derived properties. *Fundamental* properties are directly and easily measured, while *derived* properties are usually obtained from fundamental ones by some kind of mathematical relationship. Properties may be either *extensive* or *intensive*. *Extensive* properties are additive. Their determination requires evaluation of the size of the entire system; examples are volume and mass. To describe the volume of a system, a standard is chosen such as a cubic centimeter, and each equivalent cubic centimeter in the entire system is added together to give the total volume. On the other hand, *intensive* properties are independent of the size of a system. These properties are well defined in each small region of the system. Examples are density, pressure, and temperature. Any and every region in the system, if measured, reports the same value as any other. Intensive properties are not additive and a dimensional analysis will usually show them to have units written as "something" per gram or mole or volume measure, etc.

We define an intensive property as "well defined in a small region," what is meant is "small with respect to the total system, yet not too small." Consider density:

$$D = \frac{m}{V} \tag{10.4}$$

where D is the density given as a ratio of mass, m, to volume, V. The average density of a system can be determined (assuming that the density is uniform throughout the system), if a volume is chosen that is small compared to the total volume but that still contains a massive number of particles. For example, if a cubic angstrom were chosen as the volume, there would be a good chance that the volume would be filled or empty based on the fluctuations of a single molecule. In such a case, the measured density would swing drastically up and down. Defining density in such a way would not lead to a thermodynamic property. The proper sampling volume can be found by considering the volume that will give an adequate number of particles, n. It can be shown by statistical methods that the actual value of n at any given time will be a number that fluctuates around n by $n \pm \sqrt{n}$. If $n = 100$, the actual value will range from 90 to 110, while if $n = 250{,}000$, the actual volume will vary by 500 (between 249,500 and 250,500), probably an undetectable difference. If a small volume must be chosen for practical reasons, there will be cases where the volume chosen will be filled and cases where it will be empty. Then, a basic theorem of statistical thermodynamics can be used which states that averaging over (infinite) time will yield the same result as a numerical average.

10.2.4 The State of a System

With a system and its boundaries, surroundings, and properties selected, the *state of the system* can be specified. When all of the properties of the system are specified, we say the system is *defined*. It must be known, from experimental study or experience with similar systems, what properties must be considered in order that the state of a system can be defined with sufficient precision. For a system to be defined, the values of the properties must not be changing, this would make it impossible to specify those variables and hence the system could not be defined. When the properties of the defined state are unvarying over time and no flow of mass or energy occurs, a state of *thermodynamic equilibrium* exists. If there is a flow of mass or energy through the system but no change in the measured properties, the system is in a *steady state*. Steady state conditions are common in biological systems. For example, a cell in a tissue may be continuously utilizing glucose and oxygen with the production of CO_2 and H_2O. These molecules are continuously being fluxed through the system and while the measured properties are unchanging the system is not at equilibrium but is in a steady state.

The task in thermodynamics is to find the minimum number of properties necessary for a complete description of a system. An *equation of state* is a mathematical relationship that relates certain properties of a system. Equations of state are found out by fitting the equation to experimental data. The most familiar equation of state to students of chemistry is the ideal gas law:

$$PV = nRT \tag{10.5}$$

where pressure (P), volume (V), amount of gas in moles (n), temperature (T) are four variables that define the state of a dilute inert gas (R is a proportionality constant equal to 0.0821 atm/K mol). In Chapter 7 this equation as well as other equations of state derived from experiment was discussed. The tremendous predictive power of thermodynamics is unveiled when these properties can be found, measured, and the relationship of the measured properties to other important but unmeasured terms can be known. Then, by knowing the state of a system in terms of a few easily measured parameters, the values of all other properties become known. The trick to thermodynamic problem solving is to find properties in terms of others and to find the rules for doing the transformations.

The properties of a system at equilibrium depend on its state at the time in question and not on what has happened in the past. Because of this, thermodynamic properties at equilibrium are called *functions of state.* For example, the volume of a system is denoted by V. The *state function* is given by the relationships codified in Eq. (10.5) and is written in functional form:

$$V = V(P, T) \tag{10.6}$$

Any infinitesimal change in volume may be denoted as dV. dV is the differential of the function of state. This means that the integral of all the infinitesimal volume changes for any process that starts in state 1, with volume V_1, and ends in state 2, with volume V_2, will lead to a total volume change that depends only on the initial and final states:

$$\text{Volume change} = \int_1^2 dV = V_2 - V_1 \tag{10.7}$$

The key to understanding the mathematics involved in state functions lies in the fact that the value $V_2 - V_1$ is absolutely independent of the path, route, speed, complexity, or any other factor chosen in going from state 1 to state 2. A differential of a state function can always be integrated in the traditional fashion taught in calculus courses. Such differentials are called *perfect, total, or exact differentials.* Differences in state functions, such as $V_2 - V_1$, are written as follows:

$$V_2 - V_1 = \Delta V \tag{10.8}$$

The delta notation (Δ) is used only for differences in the values of state functions.

10.2.5 How Many Properties Are Required to Define the State of a System?

Just how many properties are required to fully specify the state of a system? The rule that answers this question is called the *phase rule* and will be discussed in more detail in Chapter 13. The answer can be stated simply, however. For a single homogeneous pure substance (water), one extensive property (mass) and two intensive properties (T, P) specify the state of the system. This means that with these three *degrees of freedom* fixed, all further properties of the system are automatically fixed, i.e., the viscosity, dielectric constant, refractive index. With the addition of each subsequent component to this minimal system, one additional intensive property for each component is required to fix the state of the system. A *component* is a constituent of the system whose concentration can be independently varied. The other aspect of the rule is consideration of the number of phases present. A phase is a distinct chemically and physically homogeneous region of a system such as a gas phase over a liquid phase in a system of water. Each additional phase requires one less intensive variable for system definition while requiring one additional extensive variable to define the mass of the phase. The small number of variables codified by the phase rule that are needed to define a system's state show the powerful abstraction that thermodynamics provides of the natural world. The state of a thimble full of water can be characterized with three variables thermodynamically while the number of variables needed for a molecular description is on the order of 10^{23}.

10.2.6 Changes in State

If a system undergoes an alteration in state by going from a specified initial state to a specified final state, the *change in state* is completely defined when the initial and the final states are specified. When any state function changes from state 1 to state 2 and returns to state 1 again, a cycle occurs which can be written as follows:

$$(V_2 - V_1) + (V_1 - V_2) = 0 \tag{10.9}$$

This can be written instead using the cyclical integral symbol:

$$\oint dX = 0 \tag{10.10}$$

Here X is a state function and $\oint$ is an integral around a closed path, one that returns to the original state. A process that returns to the original state is called a *cycle*. The cyclic integral of a change in state from an initial to a final state and back to the initial state is always equal to zero.

Notice that so far there has been absolutely no discussion of the incremental steps along the way from state 1 to state 2. As long as only state functions are being described, this is not a problem. However, the method of getting from one state to another is very important in certain situations. The *path* of the change in state is defined by giving the initial state, the sequence of the intermediate states arranged in the order traversed by the system, and the final state. A *process* is the method by which a change in state is effected. The description of a process consists in stating some or all of the following: (1) the boundary; (2) the change in state, the path followed, or the effects produced in the system during each stage of the process; and (3) the effects produced in the surroundings during each stage of the process.

10.3 The First Law States that “The Energy of the Universe Is Conserved”

A thermodynamic system is characterized in terms of its internal energy. In a mechanically conservative system this energy is equal to the sum of the potential and kinetic energies contained in the system (U). The internal energy of a thermodynamic system is conserved and if no external forces act on the system the internal energy remains constant. The kinetic and field energies, such as electromagnetic and gravitational fields with energy densities, all contribute to the internal energy of a thermodynamic system, e.g., a container filled with only infrared radiation (the description of a blackbody source) is a thermodynamic system. Energy can enter the system:

(1) as dissipative work that raises the temperature but is not stored in the system as increased potential energy. For example, if work is done on the system by the turning of a paddle as in Joule's experiment, we will find that the work is not stored as potential energy yet it is not lost because it has altered the thermodynamic state. This energy is found in the increased portion of energy associated with disorganized small-scale motions leading to an increased temperature. The energy is unrecoverable but not lost.

(2) via the flow of heat. Heat is the general term we use for any energy that is transferred from one system to another when the systems are placed in thermal contact. Heat can be vibrational energy, electromagnetic energy, or any other spontaneously flowing energy that flows between systems independent of the actions of the observer.

(3) by reversible work that changes property values that represent stored potential energy in the system.

A key point regarding the internal energy of a system is that once the internal energy is fixed, so are the observables of the system.

The change in the internal energy of a system is dependent on the work done, both reversible and dissipative, and the heat that passes between two systems. Once a system is defined, the *first law* dictates that the energy of the system and its surroundings will remain constant. Because the internal energy is a state function its characterization depends only on the initial and final states of the system. The first law says that the total energy of the system and its surroundings remains constant, so it would be expected that some energy may be exchanged between them. This exchange occurs at the boundary of the system and will take the form of either thermal energy (heat) or work. Intuitive and mathematical appreciation of this last statement is critical to an understanding of thermodynamics. Remember that what is being sought is a set of measurable indices that will provide information about the energy of a system. Thermal energy and work are generally measurable in the surroundings where the experimenter is located. Heat and work appear at the boundary, not in the system or in the surroundings. Indeed, heat and work are either done by the system on the surroundings or by the surroundings on the system. Energy is exchanged in this way, and the effects of heat and work can be measured in the system or the surroundings, though the heat and work themselves are really only boundary characters. It is now reasonable to say that the change in internal energy, U, of a system may be related to the work, w, and heat, q, passed between the system and the surroundings. For an infinitesimal change in internal energy, we can write the *first law of thermodynamics*:

$$dU = \text{đ}q + \text{đ}w \tag{10.11}$$

We will use the đ notation to represent the infinitesimal change in heat or work. This notation is a little archaic but will serve to remind us of an important distinction between functions involving work and heat versus energy. Heat and work are boundary entities and thus are not described by state functions but instead by *path* *functions*. By convention, both heat and work are recognized by observing changes in the surroundings and not in the system. This last point helps to highlight a very important difference between heat and work. *Work* is defined as the displacement of an object acted on by a force, thus usually converting energy into some useful form of potential energy. Traditionally, this is defined as the raising (or falling) of a weight in the surroundings. Such an action leads to a rise (or fall) in the potential energy of the surroundings and hence represents an exchange of usable energy. *Heat*, on the other hand, manifests itself as energy transfer that results in a rise (or fall) in temperature of the surroundings but that is not recoverable as useful or usable energy. This statement should not be interpreted to suggest that heat is valueless, a fact to which anyone near a fireplace on a cold night can attest.

Equation (10.11) indicates the algebraic behavior of heat and work. Since the first law dictates that the total energy of a system and its surroundings always remains constant, the transfer of energy that occurs across the boundary is either as heat or work. Since the sum of the energy in the system and the heat or work gained or lost by the system is always the same, the quantities of heat and work are treated as algebraic terms. By convention, heat that increases the temperature of the surroundings is given a negative sign (an exothermic event), while heat that raises the temperature of the system is given a positive sign (an endothermic event). Work done on the surroundings (e.g., lifting a weight in the surroundings) is called negative while work done by the surroundings on the system is said to be destroyed in the surroundings and is given a positive sign. The convention of giving work done on the system a positive sign is the modern convention. Older texts and papers will often use the older convention where work done on the system was given a negative sign. In these older writings, the first law is written as $\Delta U = q - w$.

10.3.1 Specialized Boundaries Are Important Tools for Defining Thermodynamic Systems

Knowledge of the boundary conditions is important since any change that occurs in a system must do so through a boundary. The several forms of boundaries described earlier can be employed to simplify heat–work relationships because they limit the available path functions. *Adiabatic walls* are boundaries across which no heat transfer may occur. This is the only limitation, however, and anything else may be passed through an adiabatic boundary. In an adiabatically bounded system, $\text{đ}q = 0$ and the first law reduces to

$$dU = \text{đ}w \tag{10.12}$$

In contrast to adiabatic boundaries, *diathermal* walls are boundaries where heat interactions are allowed to occur. Remember that the concept of an *isolated system* is one in which the boundary is adiabatic and work interactions (and therefore material transfer) are limited. For an isolated system, $q = 0$, $w = 0$, and ΔU=0.

The differential notation for work and heat has been introduced. The first law can be written in terms of infinitesimal change as $dU = đq + đw$. It is important to recognize that this differential notation does not represent an equivalence between U and w or q. It has already been stated that energy is a state property and work is a path function. That relationship is codified in the notation dU versus $đw$. dU is a differential that when integrated gives a perfect or definite integral:

$$\int_{\text{initial}}^{\text{final}} dU = \Delta U = U_{\text{final}} - U_{\text{initial}} \tag{10.13}$$

a quantity that is independent of the path of integration. On the other hand, $đw$ or $đq$ are inexact differentials whose integration gives a total quantity that will depend completely on the path of integration:

$$\oint đw \neq \Delta w \neq w_{\text{final}} - w_{\text{initial}} \tag{10.14}$$

$$\oint đq \neq \Delta q \neq q_{\text{final}} - q_{\text{initial}} \tag{10.15}$$

The functions defined by an inexact integral depend not on the initial and final states but rather on the path between the states. The term given for these integrals is *line integrals*. For this reason, line integrals are the appropriate mathematical expression for the path-dependent functions of both heat and work. Path functions may take any value greater than a certain minimum. The calculation of this minimum is an important aspect of thermodynamic methodology because these minima help describe reversible processes. Reversible paths are an important topic that will be covered later.

Work done on the surroundings may take many forms, all of which are equivalent because of their shared lineage in Eq. (10.1). For example, in many cases (especially in physical chemistry textbooks) the work most discussed is the mechanical work that results from the expansion of a volume and that is described by pressure and volume changes, i.e., $–P_{\text{ext}}dV$. Many other forms of work can be considered and all are treated in the same fashion. Important representations of work in biological systems include mechanical work related to moving a weight in a gravitational field, $mg\ dh$; electrical work resulting in charge transfer, ψdQ; and surface area changes, γdA. A system in which all of the above forms of work could be done can be described as

$$đw = mg\ dh - P_{\text{ext}}dV + \psi dQ + \gamma dA + \cdots \tag{10.16}$$

where m is mass, g is the acceleration of gravity, h is height, ψ is electrical potential, Q is charge, γ is the surface tension, and A is the surface area. If the only concern is a single kind of work, say, pressure–volume work, then the expression can be simplified:

$$đw = -P_{\text{ext}}dV \tag{10.17}$$

Using the example of pressure–volume work, an illustration of a line integral can be made. Consider two states of a system, state 1 defined by P_1 and V_1 and state 2 defined by P_2 and V_2. The problem is to find the compression work done on the system and hence destroyed in the surroundings in going from state 1 to state 2. In Fig. 10.1, the work done is represented as a path between state 1 and state 2. In both cases shown, the initial and final states are the same. However, the path taken to get from 1 to 2 differs between Fig. 10.1a and b. Consequently, the work, which is the integral of the path and is represented by the shaded area of the graph, is dependent on the path and not on the initial or final states. It should be obvious that a specific minimum value of work must be destroyed in the surroundings for the compression to be performed, but that virtually no limit to the amount of work destroyed in the surroundings exists if the path is appropriately chosen. For a cyclic function of a path, $\oint w$, the value may be zero or any other value at all, while for all state functions $\oint X = 0$. These path properties described for work are valid, in addition, for the behavior of heat.

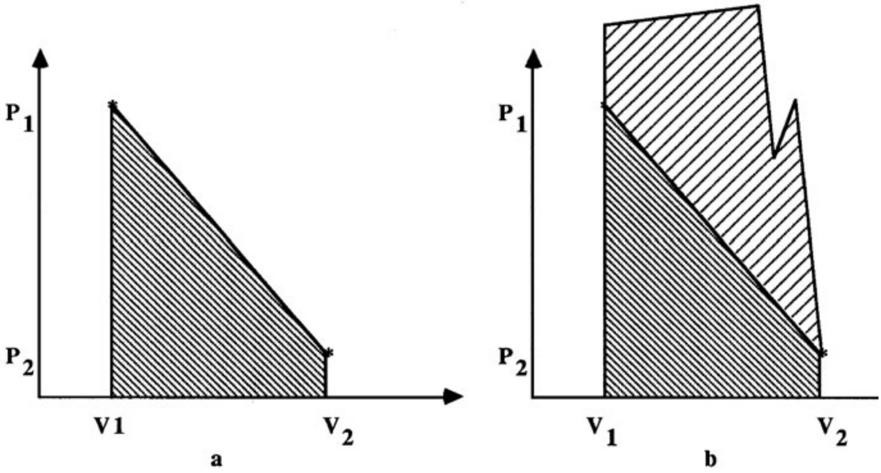


Fig. 10.1 Work, represented by the area under a curve, depends on the path and not on the initial and final coordinates, P_1, V_1, and P_2 V_2

10.3.2 Evaluating the Energy of a System Requires Measuring Work and Heat Transfer

A fundamental goal of thermodynamics is to quantitate and describe the internal energy of a system. The first law can lead to methods to determine the energy of a system. Knowing this internal energy is valuable in the study of biochemistry and cellular systems, since life processes are highly energy dependent. Boundary limitations are employed to simplify and define the relationship between the work and the state of a system. Consider a system with adiabatic boundaries: only work appears at the boundary. An infinitesimal amount of work is done on or by the surroundings, $\text{đ}w$. Since $q = 0$, then

$$dU = \text{đ}w \tag{10.18}$$

Integrating gives

$$\Delta U = \int dU = \int \text{đ}w = w \tag{10.19}$$

In an adiabatic system, the only way that energy can be exchanged is through work. Since thermal interactions are not allowed, any work done on the system will directly change the related energy values; for example, if an object is subjected to an accelerating force, a new velocity will be attained without loss due to friction and heat production. In a diathermal thermodynamic system, work will not necessarily have the same effect. Here the object may be subjected to an accelerating force, but friction may prevent any change in velocity at all. The entire work destroyed in the surroundings may appear as a change in thermal energy at the boundary. In a diathermal system, then, knowing the work expended does not necessarily lead to knowledge of the change in the system's energy.

ΔU can, however, be determined experimentally in systems that are diathermal. If an adiabatic path can be found to exist between two states, then Eq. (10.19) becomes valid for all systems. It can generally be written for this situation:

$$dU = \text{đ}w_{\text{adiabatic}} \tag{10.20}$$

This will provide a technique for determining ΔU. It is not possible to determine the absolute energy of a system by evaluating the path function for work. However, since it is the change in energy between states that is usually the important issue, this is adequate and very practical information.

It is not always convenient to find an adiabatic path for a given change in state. Does this mean that knowledge about energy transformation is not available? Fortunately, the first law provides a solution to finding ΔU even for a diathermal system in which both heat and work are exchanged. The derivation follows. Consider: $\oint dX = 0$ for any state function, while $w_{\text{cyclic}} = \oint \text{đ}w \neq 0$ and $w_{\text{cyclic}} = \oint \text{đ}q \neq 0$. The first law states that in a cyclic transformation, the work produced in the surroundings $(-\text{đ}w)$ is equal to the heat withdrawn from the surroundings $(+\text{đ}q)$. Consequently

$$\oint (-\text{đ}w) = \oint \text{đ}q \tag{10.21}$$

This is equivalent to

$$\oint (\text{đ}q) + (\text{đ}w) = 0 \tag{10.22}$$

and

$$\oint (\text{đ}q + \text{đ}w) = 0 \tag{10.23}$$

Because any cyclical function that is equal to zero must also be a state function, this equation may be defined as a state variable U, the energy of the system. Therefore

$$dU = \text{đ}q + \text{đ}w \tag{10.24}$$

Integrating, using line (i.e., path) integrals where appropriate

$$\int dU = \int \text{đ}q + \int \text{đ}w \tag{10.25}$$

gives

$$\Delta U = q + w \tag{10.26}$$

Therefore, the first law now provides a solution for finding ΔU, even if no convenient adiabatic path exists. By measuring the heat, q, and the work, w, of a path, the change in the state variable of internal energy may be found.

It is worth restating that the value of thermodynamics lies in its ability to provide information about systems and conditions that are unfamiliar and previously unseen. This power lies in the quality of state functions that constrain a real system so that by measuring just several of a system′s state variables, other state variables, including the one of interest, can be determined. In the most familiar case, the equation of state for an ideal gas (see Eq. (10.5)) indicates that if only P and V are known for a system, then T can be determined because of the natural constraints of the system. Likewise, if T and V are known, P can also be found for any corresponding system. Once ΔU is determined, the state of the system under study can be described in terms of other state variables. This is a very valuable feature. The minimum number of state variables necessary to completely describe a particular system is given by a simple formula, the phase rule.

Consider describing a system of energy, U (found by a path method as described earlier), in terms of other state variables, for example, temperature, T, and volume, V. It follows from the discussion above that since U is a state variable as are temperature and volume, then a change in U, the energy of the system, could be described in terms of T and V. Energy is therefore a function of the temperature and volume of a system:

$$U = U\ (T, V) \tag{10.27}$$

Equation (10.27) is a simplification. In fact, U, or any other state function, is a function of all other constrained values in the system under study. In other words, each state variable depends to some degree on every other variable (and there is virtually an infinite number) in a given system. To be entirely accurate in describing a thermodynamic system, it would be necessary to elucidate each and every state variable:

$$U = (U\,T,\ V,\ P,\ X_i,\ S,\ G, \ldots) \tag{10.28}$$

The symbols for mole number (X_i), entropy (S), and free energy (G) are introduced here and will be discussed subsequently. Defining all these choices would be a prodigious and surely frustrating task. However, in the real world such exhaustive degrees of accuracy are not necessary. Therefore, scientists are saved from the drudgery of such complete descriptions. Reasonable simplifications are almost always preferred. By using partial derivatives, the state function of interest can be written and examined in terms of the other variables:

$$dU = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV \tag{10.29}$$

Expressions like the one above allow the evaluation of several variables at once. This analysis is made possible by differentiating just one variable while holding all the other variables constant. This process is repeated with each variable, until all are differentiated. The result is a differential equation that mathematically relates each of the state variables to each of the others in a simultaneous manner. One of the most important caveats that accompanies the use of partial differential statements is the need to always interpret these expressions in terms of their physical as well as mathematical meaning. Such awareness is necessary to ensure their effective use as well as to prevent the frustrating confusion that can be associated with the use of these equations. The point in using partial differential expressions is to find mathematically accurate expressions that can be manipulated but that are measurable in a real system.

10.4 The Heat Capacity Is a Property that Can Reflect the Internal Energy of a System

The thermodynamic exercise now remains to discover what property (or properties) of the system can be measured that will allow the determination of ΔU. There are two equations discussed so far that describe dU:

$$dU = \text{đ}q + \text{đ}w \tag{10.30}$$

and

$$dU = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV \tag{10.31}$$

These are equated, giving

$$\text{đ}q + \text{đ}w = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV \tag{10.32}$$

For the system under discussion, only volume change work is possible, so $\text{đ}w$ can be written as

$$\text{đ}w = -P_{\text{ext}} dV \tag{10.33}$$

Substitution gives

$$\text{đ}q - P_{\text{ext}} dV = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV \tag{10.34}$$

Now we use the principle of judiciously constraining the system under study. If the task is to find the change in energy of a system that is kept at constant volume, then dV becomes zero and the equations from above can be rewritten as follows:

$$dU = \text{đ}q_v \tag{10.35}$$

and

$$dU = \left(\frac{\partial U}{\partial T}\right)_V dT \tag{10.36}$$

Again equating these identities gives the following result:

$$\text{đ}q_v = \left(\frac{\partial U}{\partial T}\right)_V dT \tag{10.37}$$

This is now an expression that relates the heat drawn from the surroundings to the increase in temperature of the system at a constant volume. Because both $\text{đ}q_v$ and dT are easily measured experimentally, the task of relating the equation to the real system is approaching completion. The ratio of these two experimentally accessible parameters

$$\frac{\text{đ}q_v}{dT} \tag{10.38}$$

is called the *heat capacity* (at constant volume) or C_v . Writing this in summary gives

$$C_v = \frac{\text{đ}q_v}{dT} = \left(\frac{\partial U}{\partial T}\right)_V \tag{10.39}$$

Several favorable results are apparent here. First, the partial derivative discussed above, $\left(\partial U / \partial T\right)_V$, has been solved. Substituting gives

$$dU = C_V dT \tag{10.40}$$

Integrating this expression will give the following result:

$$\Delta U = \int_{T_1}^{T_2} C_V dT = C_V \Delta T \tag{10.41}$$

It is now possible to calculate ΔU directly from the properties of a system if the heat capacities are known. Extensive tables of heat capacities are available for many materials. C_v is always positive, and therefore whenever heat is added to a system, the temperature of the system and the system's internal energy will rise. For a system at constant volume constrained to do only *PV* work, temperature therefore is a direct reflection of the internal energy.

Heat capacity is expressed in units of joules per degree mole in the SI system. Much literature still uses the units of calories per degree mole, however. The heat capacity relates on a molar basis just how much heat it takes to raise the temperature of a system by one degree Celsius or Kelvin. Heat capacities often vary with temperature and depend upon the volume and pressure. In the example above, the derivation of heat capacity was for a system held at constant volume. In a system that is held at constant pressure, the heat capacity is expressed as C_p. C_v is just about equal to C_p for liquids and solids. For a gas, heat capacities under these differing conditions vary by a factor of *R*, the individual gas constant for a specific gas:

$$C_p = C_v + R \quad \text{[for a gas]} \tag{10.42}$$

$$C_p \approx C_v \quad \text{[liquids and solids]} \tag{10.43}$$

10.5 Enthalpy Is Defined When a System Is Held at Constant Pressure

Just as C_v was derived by considering a system at constant volume, the constraints on a system can be and should be guided by the practical aspects of a specific system under study. Defining systems at constant pressure is realistic for most biochemists and life scientists, because most of the real systems studied in the organism or in vitro are constant pressure systems. An analysis of the first law under this constant pressure constraint will lead to a new state function, *enthalpy, H*. Remember that the goal is to find an experimentally measurable value that can inform on the internal energy state of the system under study, given the constraints chosen.

Starting again with the first law:

$$dU = \text{đ}q + \text{đ}w \tag{10.44}$$

and examining a system that does only pressure–volume work, so $\text{đ}w = -PdV$, now

$$dU = \text{đ}q - PdV \tag{10.45}$$

Pressure is a constant in this system and so integration can be immediately carried out:

$$\int_1^2 dU = \int_1^2 \text{đ}q_p - \int_1^2 PdV \tag{10.46}$$

which gives

$$U_2 - U_1 = q_p - P\,(V_2 - V_1) \tag{10.47}$$

By algebra and by substitution of the identities $P_1 = P_2 = P$ at constant pressure:

$$(U_2 + P_2V_2) - (U_1 + P_1V_1) = q_p \tag{10.48}$$

Now both P and V are already defined as state functions and a product of state functions is itself dependent only on the state of the system. Consequently, the expression:

$$U + PV \tag{10.49}$$

is a state function. This new state function is given the symbol H and the name, *enthalpy:*

$$H = U + PV \tag{10.50}$$

Substituting H for $U + PV$ in Eq. (10.48) gives the following:

$$(H_2 - H_1) = q_p = \Delta H \tag{10.51}$$

Having the result, that the heat withdrawn from the surroundings is equal to the increased enthalpy of the system, is especially valuable if the energy change, ΔU, is now considered with a real simplification of most aqueous phase and biological processes, $\Delta V = 0$. Consider that

$$\Delta U + P\Delta V = q_p \tag{10.52}$$

and with $\Delta V = 0$, a reasonable approximation can be written that

$$\Delta U = q_p \tag{10.53}$$

and since

$$q_p = \Delta H \tag{10.54}$$

then

$$\Delta U = \Delta H = q_p \tag{10.55}$$

Enthalpy is a valuable state function because it provides a method for determining a realistically constrained biological or aqueous phase system′s energy simply by measuring the heat exchanged with the surroundings. By writing enthalpy in terms of temperature and pressure, $H = H(T, P)$, in a fashion similar to that done earlier for C_v, it can be shown that the heat capacity at constant pressure, C_p, is related to enthalpy as follows:

$$\Delta H = C_p \Delta T \tag{10.56}$$

The actual derivation of this is left to the reader as an exercise.

Enthalpy is a useful state function to the biochemist for a variety of practical reasons. First, most chemical and biological processes of interest occur at constant pressure and with no change in volume. Hence the internal energy of a system is

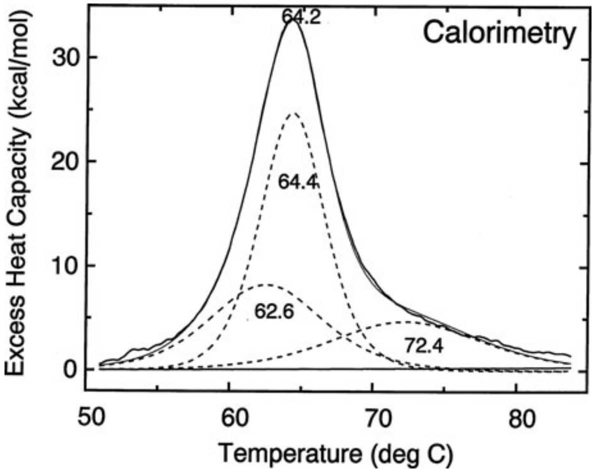


Fig. 10.2 High-resolution calorimetry of β_2 glycoprotein I. By measuring the temperature change as heat is added to a system, the changes in heat capacity can be recorded. Since the heat capacity is a function of the internal energy states of the system it can be sensitive to changes in the structure and organization of the system. Here the *solid lines* show the raw and theoretical excess heat capacity curves for the reversible unfolding of β_2 glycoprotein I in an aqueous buffer at physiological pH. The *dashed lines* show the maximum number of reversible, two-state transitions which are found following deconvolution analysis of the experimental curve. There are three reversible states. Thus this calorimetric study suggests a minimum of three independently folded domains in this glycoprotein. (Courtesy of Drs. Bin Lu and Mary T. Walsh)

Table 10.1 Standard thermodynamic properties of substances (298.15 K). Values in KJ/mol

Compound	$\Delta H_f°$	$S°$	$\Delta G_f°$
Inorganic			
Ag	0.00	42.55	0.00
Ag^+(aq)	105.58	72.68	77.107
C(g)	716.68	158.10	671.26
Cl_2(g)	0.00	223.01	0.00
Cl^-(aq)	−167.16	56.5	−131.23
H_2(g)	0.00	130.69	0
H_2O(g)	−241.82	188.825	−228.572
H_2O(l)	−285.83	69.91	−237.13
H_2S(g)	−20.63	205.79	−33.56
H^+(aq)	0.00	0.00	0.00
NH_3(g)	−46.11	192.45	−16.45
NH_3(l)	−80.29	111.3	−26.50
NaCl(s)	−411.15	72.13	−384.14
NaCl(aq)	−407.27	115.5	−393.13
Na^+(aq)	−240.12	59.00	−261.91
O_2	0.00	205.14	0.00
O_3	142.7	238.93	163.2
OH^-(aq)	−229.99	−10.75	−157.24
Organic			
Methane	−74.81	186.264	−50.72
Ethane	−84.68	229.60	−32.82
Methanol	−238.57	126.80	−166.23
Ethanol	−276.98	160.67	−174.14
Acetic acid	−484.10	159.83	−389.36
Benzene(g)	82.93	269.20	129.66
Benzene(l)	49.04	173.26	124.35
Cyclohexane(g)	−123.14	298.24	31.76
n-Heptane(g)	−187.78	427.90	7.99
n-Octane(g)	−208.45	466.73	16.40
Naphthalene(g)	150.58	333.10	224.10
Naphthalene(l)	78.53	167.39	201.585
Adenine	95.98	151.00	299.49
L-Alanine	−562.70	129.20	−370.24
L-Glycine	−537.25	189.95	−490.57
L-Tyrosine	−671.5	214.01	−197.15
L-Glutamic acid	−1009.68	118.20	−731.28

Source: Stull D.R., Westrum E.F. and Sinke G.C. (1969) *The Chemical Thermodynamics of Organic Compounds*. Wiley, New York; The NBS Tables of Chemical Thermodynamic Properties (1982) Wagman D.D. et al. (eds.) *J. Phys. Chem. Ref. Data*, **11**:Suppl. 2.

easily and quite accurately approximated by measuring the enthalpy. Second, given these constraints, the experimental measurement of q_p is both easy and reasonably precise. Finally, enthalpy is easily predicted from a state property, C_p, thus allowing inference of changes in enthalpy for systems never seen before. As indicated in the introduction to this section, this is precisely the valuable kind of tool that thermodynamics can provide the practical scientist.

A great deal of information based on enthalpy is available, including tables listing heats of formation, fusion, solution, dilution, reaction. These enthalpic values are valuable precisely because they indicate the change in energy state of the system under study. By measuring the energy changes associated with changes in molecules, a large amount of structural and functional information can be deduced. Calorimetry is an important technique that measures changes in heat capacity and thus reflects the behavior and properties of the system (often a molecule) under study (Fig. 10.2). Most readers of this volume will already be familiar with the use of enthalpy to determine heats of reaction and formation of compounds as well as heats associated with phase changes and solvation. Most texts in introductory chemistry and physical chemistry provide excellent coverage of the practical use of this state function including the use of Hess's law, standard states, and calculation of heats of formation, phase change. The thermodynamic properties for a variety of molecules of biological interest are listed in Table 10.1.

Enthalpy will play an important role because of its ability to provide a good picture of the energy changes associated with a process. But enthalpy and the first law alone will not adequately predict the spontaneous direction of a reaction. However, when combined with the state function entropy, a function that represents the second law, a new state function, the *Gibbs free energy*, is the result. The Gibbs free energy indicates the spontaneous tendency of a system to move in one direction or another. Enthalpy will reappear therefore after an examination of the second law of thermodynamics.

Thought Questions

1. Make a diagram summarizing isolated, closed, and open systems in terms of their boundary properties.
2. Comment on the following statement: Like the wave–particle duality, the thermodynamic–statistical mechanical duality forms an epistemological boundary to our knowledge of a system.

Further Reading

Most physical chemistry texts treat thermodynamics in moderate detail. All of the texts listed in Chapter 2 cover the topics. Thermodynamics is a field of physics used by chemists and biologists. For a text devoted to chemical thermodynamics:

Klotz I.M. and Rosenberg R.M. (1994) *Chemical Thermodynamics, Basic Theory and Methods*, 5th edition. Wiley, New York.

And for a treatment of biological thermodynamics:

Haynie D.T. (2001) *Biological Thermodynamics*. Cambridge University Press, Cambridge.

Two small monographs written by physicists are

Fermi E. (1936) *Thermodynamics*. Dover, New York.

Schrödinger E. (1989) *Statistical Thermodynamics*. Dover, New York. (Reprint of the 1946 lectures given to the Dublin Institute for Advanced Studies.)

A fresh approach drawing from quantum mechanics is

Waldram J.R. (1985) *The Theory of Thermodynamics*. Cambridge University Press, Cambridge.

For the specialized problem of the small system which is important for the biologist:

Hill T.L. (1994) *Thermodynamics of Small Systems, Parts I and II*. Dover, New York.

Applications of Calorimetry

Chaires J.B. (2008) Calorimetry and thermodynamics in drug design. *Annu. Rev. Biophys.*, **37**: 135–151.

Problem Sets

1. For each of the following, identify the boundary, surroundings, and system. (a) a cell, (b) a flask in an incubator, (c) the world, (d) a leaf, (e) the brain, (f) the known universe, (g) the Earth, (h) a sunflower seed.
2. Perform a systems analysis of a general thermodynamic system. Make a parallel analysis of each of the thermodynamic systems above in terms of their properties, elements, relationship rules, and background/contextual space.
3. For the following properties of a system related to matter identify which are intensive and which are extensive? (density, concentration, mass)
4. Which of the following properties related to PVT are intensive and which are extensive? (specific volume, molar volume, pressure, temperature, volume)
5. Which of the following properties related to thermal energy are intensive or extensive (heat capacity, chemical potential, molar energy, entropy, free energy, specific heat (C_p/gram), enthalpy, molar entropy, molar enthalpy, energy)?
6. Which of the following property used to characterize a system are intensive or extensive (dielectric constant, volume, moles of solvent, mass, refractive index, temperature, viscosity, zeta potential)?
7. The energy of a system is easily measured by thermal techniques. When the state of a system and a change in the state observables is temperature dependent, an exothermic or endothermic change can be measured. For the following list of events indicate whether a physical change is endo- or exothermic.

Adsorption
Desorption
Freezing
Melting
Chemisorption
Vaporization
Sublimation
Dehydration
Desolvation

8. What amount of heat is needed to change the temperature of 0.5 l of H_2O from 275 to 325 K? The pressure is 1 atm. $C_p = 1$ cal/g deg?
9. If the same heat were added to an equivalent mass of ethanol ($C_p = 111.4$ J/K mol) what would the temperature change of the liquid ethanol be?

10. Using the bond energies given below calculate the heat of formation of gaseous hexane, cyclohexane, and benzene. Compare your answers with the thermodynamic heats of formation given in Table 10.1. Explain any discrepancies.

Bond type	Bond dissociation energy (kJ/mol)
C–C	344
C (graphite)	716
C–H	415
C=C	615
H2	436
O2	498

11. One mole of water at 10°C at 1 atm is added to 1 mol of water at 30°C. What is the final temperature? No heat leaves the system.
12. (a) The ideal temperature for a certain refreshment is 4°C. How much ice, cooled to –20°C should be added to 250 ml of the refreshment at 25°C in order to cool it to 4°C with no dilution of the drink (i.e., no ice melts into the drink)? (b) If the drink is served in a 300 ml cup (maximum ice = 50 cc), what is the minimal amount of ice that can be added to achieve temperature and have minimal dilution? Assume the system to be adiabatic. The molar *Cp* of ice is 37.8 J/K mol.

Chapter 11
Entropy and the Second Law of Thermodynamics

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_11,

11.1 The Arrow of Time and Impossible Existence of Perpetual Motion Machines Are Both Manifestations of the Second Law of Thermodynamics

It is a common experience that watches run down, bouncing balls stop bouncing, and even the most perfectly designed and crafted engine will eventually cease operating. Time marches on and all things, left to themselves, stop. This universal tendency for all objects and processes to move "with the arrow of time" in the direction of running down is the hallmark of the second law of thermodynamics. The obverse to this relentless natural tendency for things to run down is the (equally relentless) search for a technical fountain of youth: the system that will produce more energy or work than it consumes. Such machines are called *perpetual motion machines* and are imagined in two varieties: perpetual motion machines of the first and second kind. The first is a machine that produces more energy than it absorbs as work or heat from the surroundings and thus creates energy. This kind of device has never been found (though people attempt it every day), and the first law of thermodynamics forbids its existence. The first law permits the second type of perpetual motion machine because it draws heat from the surroundings and converts that heat into work without changing the state of the surroundings. However, it turns out that there are substantial limitations on the construction of a machine or system that can take heat and convert it into directed energy or useful work. Were this not so, it would be possible to build a machine that

1. Draws heat from the surroundings (there is an almost limitless supply of random thermal motion).
2. Converts the heat to useful work.
3. In the course of using the directed energy to do work, returns the energy to the surroundings as dissipated or scattered energy (i.e., as random thermal fluctuations).

This much more sophisticated type of perpetual motion machine is forbidden by the second law of thermodynamics, which codifies the constraints placed on building real machines capable of converting heat to work. Proposing this second kind of perpetual motion machine is an error sometimes made by biologists who view living systems as "special" since they seem so replete with vital energies that are not found in the inanimate surroundings. The conclusion is occasionally made that biological systems are unique sorts of perpetual motion machines. The problem with this analysis and the concept of these machines of the second type is that they invariably change the state of the surroundings; whereas the system seems to be full of a never-ending source of energy, the rest of the Universe is running down.

There are multiple possible statements of the second law. Some of the more useful paraphrases are included in the following list; of these, the first two expressions are the more traditional representations of the second law:

(1) No process is possible where heat is transferred from a colder to a hotter body without causing a change in some other part of the Universe.
(2) No process is possible in which the sole result is the absorption of heat (from another body) with complete conversion into work.
(3) A system will always move in the direction that maximizes choice.
(4) Systems tend toward greater disorder.
(5) The macroscopic properties of an isolated system eventually assume constant values.

11.1.1 The Movement of a System Toward Equilibrium Is the Natural Direction

Definition (#5) is useful because it leads naturally toward a definition of *equilibrium*. When the state values of an isolated system no longer change with time, the system may be said to be at equilibrium. Movement of a system toward equilibrium is considered the "natural" direction.

A state function that could indicate whether a system is moving as predicted toward a set of unchanging state variables would be of great value. Such a state function exists and is the state function of the second law, *entropy*, and it is given the symbol ***S***. Originally the derivation of S was accomplished by evaluating a Carnot cycle. We will briefly review the Carnot cycle because its powerful conclusions, especially given its origins as a Gedanken experiment intended to answer a practical question, are instructive.

Following this discussion, we will explore the elements of the statistical thermodynamic approach and will discover that, in terms of entropy, we are led to the same mathematical description as the one derived from the Carnot cycle. In the statistical approach, we say that a system will tend to maximize choice. Sometimes, this statement of the second law is written as "systems tend toward increased disorder" or "systems tend toward increased randomness." These paraphrases of the second law are useful, but care should be taken with their use because the terms disorder and randomness often lead to confusion. Thinking of the natural direction of a system in terms of its ultimate uniformity or choice is ultimately the intuitive as well as practical treatment of entropy.

11.2 The Design of a Perfect Heat Engine Is an Important Thought Experiment

Historically the fundamental problem of how to design an efficient engine that converts heat into work was the starting point for the development of thermodynamics. The engineer Sarni Carnot devised an approach that uses a *reversible* cyclic path to explore this question. This treatment is known as the Carnot cycle. We emphasize that the concept of *reversibility* is not limited to a Carnot cycle and that *the ideas developed in the following section are completely general*.

11.2.1 Reversible Paths Have Unique Properties Compared to Irreversible Paths

The ideas of reversible and irreversible paths will be developed here for a system in which pressure–volume work is being done. This choice is simply one of convenience and is practical for the heat engine under consideration. In a *reversible process*, the steps of the path occur in a series of infinitesimal steps. The direction of the path may be reversed by an infinitesimal change in the external conditions prevailing at each step.

Consider an isothermal expansion in which a piston encloses a system of a gas with temperature T_1, pressure P_1, and volume V_1. On top of the piston is a weight of mass M. The mass is part of the surroundings, and if it is moved upward, work will be done by the system on the surroundings. Conversely, if the mass pushes the piston head down, work will be done on the system. The container enclosing this system is specially designed for this example and has a series of stops that may be inserted through the side of the container (Fig. 11.1). The power stroke, that is, the raising of the weight in the surroundings, is complete when the piston head reaches the top set of stops. At this point, the state of the system will be defined as V_f, P_f, T_1. When the system is in the initial state, the volume occupied by the gas is smaller than in any other state and the pressure of the gas is higher than in any other state (each state represented by a set of stops). As long as the internal pressure of the gas is greater than the external pressure, as represented by the weight on the piston, the gas will expand and the piston will be driven upward. Since the internal pressure will be greater closer to the initial state due to the higher compression of the gas, the external pressure on the piston can be set higher at the early stages of the expansion. By continually decreasing the weight on the piston as the gas inside expands, the power stroke of the piston can be sustained until the final state is achieved.

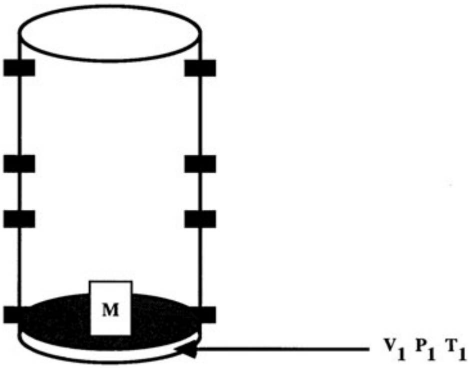


Fig. 11.1 Special container for reversible isothermal expansions and compressions

The following experiments are performed with this device. Initially, two sets of stops are inserted (see Fig. 11.2). The first set maintains the position of the piston head at the initial state of V_1, P_1, T_1. The second set of stops restrains the system to

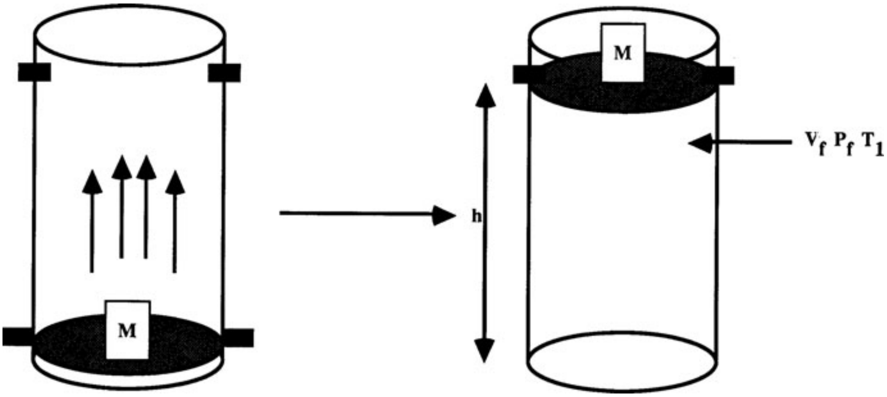


Fig. 11.2 Single-step expansion where $w = -P_{\text{ext}}\Delta V$

the state described by V_f, P_f, T_1. Since this two-stop system will allow motion from the initial to the final step, the maximum weight that can be placed on the piston head must exert a pressure just slightly less than the final pressure, P_f. Note that this weight is significantly smaller than the weight (or P_{ext}) that could be moved against by the initial pressure, P_1. Now the lower set of stops is removed, and the piston moves upward, halting at the second set of stops. The new state of the system is now V_f, P_f, T_1. The expansion to the new state has resulted in the raising of the mass, m, in a gravitational field. The work done is mgh, and because pressure is simply force over an area, A (the area of the piston), and the external pressure (including the mass, m) is the force opposing the movement of the piston, the following can be written:

$$\frac{mg}{A} = P_{\text{ext}} \tag{11.1}$$

$$w = mgh = P_{\text{ext}}Ah \tag{11.2}$$

Ah is the volume change $V_2 - V_1$, and so this is a restatement of the problem derived earlier in the section on pressure–volume work, and the resulting equation holds:

$$w = -P_{\text{ext}}\,\Delta V \tag{11.3}$$

This is the correct mathematical statement for all real expansions as long as the external or opposing pressure, in this example provided by the mass on the piston, remains the same throughout the expansion. Now we repeat the experiment with a third set of stops in place (Fig. 11.3). A weight is chosen that will allow expansion to the state V_2, P_2, T_1. Because this intermediate state has a pressure P_2 that is greater than P_f, the weight chosen for the initial expansion is larger than the weight

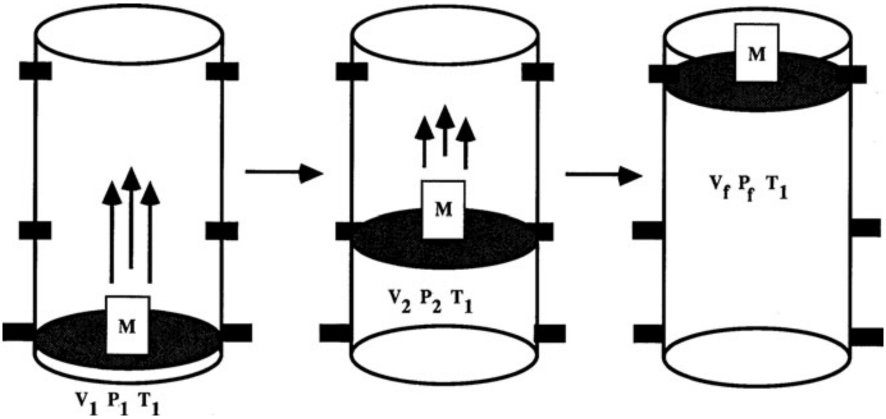


Fig. 11.3 Two-step expansion where weights are changed at each step of the expansion

chosen for the one-step expansion. Therefore, for the distance traveled, more work is done in the first part of the expansion than occurred in the one-step case. Now, the piston is allowed to move to the position described by V_2, P_2, T_1. The weight is now changed to allow the full expansion to the final state. Because the final internal pressure will be lower at V_f, P_f, T_1, the weight must be reduced before the two-step expansion is completed. After the weight is changed, the piston is allowed to move to the final position with state functions V_f, P_f, T_1. As can be seen in Fig. 11.4, the work done on the surroundings that results from the two-step expansion is greater than that resulting from the one-step expansion, even though in both cases the final state is the same.

If an infinite series of infinitesimal changes in the movement of the piston are made, each accompanied by the continuous readjustment of the weight on the piston (and hence P_ext), the maximum work available from the expansion will result. Two points are evident. First, an expansion of an infinite number of steps would take an infinite amount of time to complete. Second, the internal equilibrium of such a process will be disturbed only infinitesimally, and in the limit no disturbance at all will take place. It is reasonable to consider therefore that this system does not depart from equilibrium, except infinitesimally, throughout the entire path. Consequently, the state variable describing a system that is undergoing a reversible process can be treated as if the system were at equilibrium.

In summary, for a reversible process, the path will take an infinite time to complete, but the maximum work will result *only* from this reversible path. Furthermore, reversible processes are essentially at equilibrium except for infinitesimal perturbations and may be treated as if they are equilibrium conditions. Because of these properties, in a reversible process the external forces, such as P_ext in the example, become uniquely related to the internal forces, that is

$$P_\mathrm{ext} = P_\mathrm{int} \pm \text{infinitesimal} \tag{11.4}$$

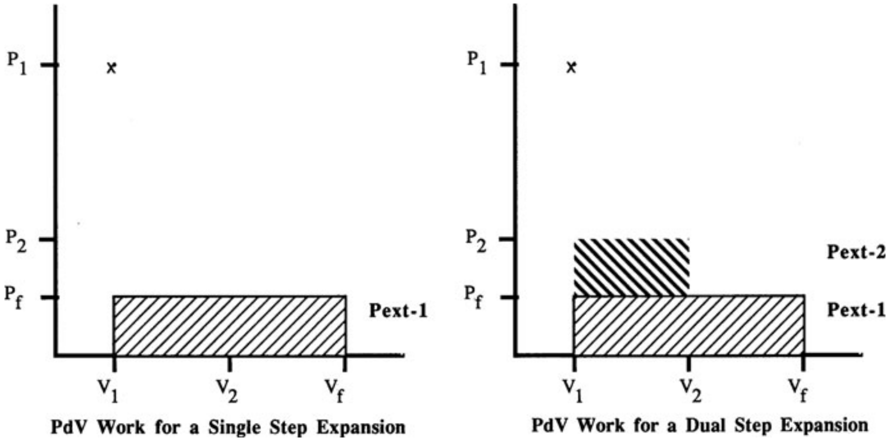


Fig. 11.4 Work function graph showing the work produced for a one- and a two-step expansion. In the one-step expansion, the work is found equal to $-P_{\text{ext}-1}\Delta V$ and, since $P_{\text{ext}-1}$ is just an infinitesimal less than P_f, the maximum amount of work available is limited. In the two-step process, a higher weight, represented by $P_{\text{ext}-2}$ is used for the initial expansion step, then the weight on the piston is reduced to $P_{\text{ext}-1}$ and the expansion is completed. More work is extracted in the two-step case because of the change in weight along the path

Thus

$$P_\text{ext} = P_\text{int} = P \tag{11.5}$$

Work for a process like this can be found in the fashion

$$đw = -PdV \tag{11.6}$$

which by integration gives

$$\int_1^2 đw = -\int_{V_1}^{V_2} PdV \tag{11.7}$$

By using the equation of state $PV = nRT$ and the identity $P_1V_1 = P_2V_2$ and substituting, the following equation is derived:

$$w = -nRT \ln \frac{P_1}{P_2} \tag{11.8}$$

This is the maximum work that is available from an isothermal expansion and therefore from a reversible path.

Where does the energy come from for this process? Recalling that this is an isothermal expansion, $\Delta T = 0$ for the entire process. Previously it was shown that

$$\Delta U = C_{\mathrm{v}} \Delta T \tag{11.9}$$

Using this equation

$$\Delta U = 0 \tag{11.10}$$

By the first law of thermodynamics

$$\Delta U = q + w \tag{11.11}$$

so

$$0 = q + w \tag{11.12}$$

and

$$w = -q \tag{11.13}$$

for an isothermal reversible expansion. Therefore, in a reversible isothermal process, there is a perfect exchange of heat from the surroundings for work done on the surroundings.

A few moments of reflection on this discussion make it obvious why the second law of thermodynamics can be stated with such clarity and certainty. The reason that all processes generate less work than the energy they consume is also explained by the nature of the perfect transforming process, the reversible path. Any path that is reversible must have an infinite series of infinitesimal steps. It will take an infinite amount of time to complete. Such a process would never be accomplished in the real world, and hence any real process may only approach reversibility. At best, the entropy of the Universe remains constant by having only reversible processes at work; but in reality no reversible process can be completed in a finite time frame, and hence, any real process leads to an increase in entropy.

Knowing that real, and therefore *irreversible*, *processes* inevitably lead to an increase in entropy is all the more valuable, if a state function can be derived that will indicate in what direction and how far a system is from equilibrium. Recall that a state function is one whose cyclic integral is equal to zero:

$$\oint X = 0 \tag{11.14}$$

Specifically, an entropy state function is required such that

$$\oint S = 0 \tag{11.15}$$

The change in entropy should be zero at equilibrium and should be positive or negative otherwise to indicate the necessary direction to reach equilibrium.

11.2.2 A Carnot Cycle Is a Reversible Path Heat Engine

A machine that is able to transform heat into work is called a heat engine. It is not possible to build a machine that extracts heat from a single heat reservoir at a uniform temperature with no other change in the system, but if two reservoirs of different temperature, T_1 and T_2 exist, a heat engine can be constructed. One process that converts heat to work is called a *Carnot cycle*. It is a reversible cycle involving two isothermal and two adiabatic paths. Such a cycle is illustrated in Fig. 11.5. The state of a fluid in two isothermal paths (*AB* and *CD*) and two adiabatic paths (*BC* and *DA*) at two temperatures is illustrated on a volume–pressure graph. The Carnot cycle is the reversible path, *ABDCA*.

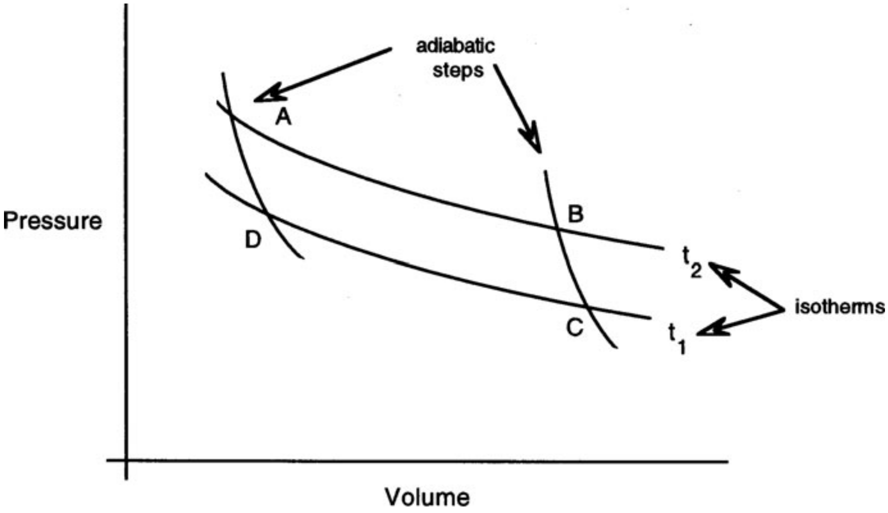


Fig. 11.5 A Carnot cycle

A heat engine is built (Fig. 11.6) with cylinder walls and piston head that do not conduct heat. The only surface capable of conducting heat is the base. The two reservoirs are so large that adding or extracting a finite amount of heat will not change their temperatures. The temperature of T_2 is greater than T_1. When we start the cycle, the state of the engine (the system) is P_A,V_A which is on the isotherm of T_2 at point *A*.

Our cycle begins. The piston is placed base down in thermal contact with reservoir T_2. Because the temperatures between the systems are the same, there is no heat flow through the base. Now an isothermal expansion is performed by slowing raising the piston until volume V_B is reached. Since the volume is increased by an isothermal process, heat passes from T_2 into the system. At the end of the expansion the system is at *B*. The cylinder is now placed on an insulating block so heat flow into or out of the system is prevented. A further expansion, this time adiabatic, is performed reversibly so that the volume further increases and the pressure and temperature of

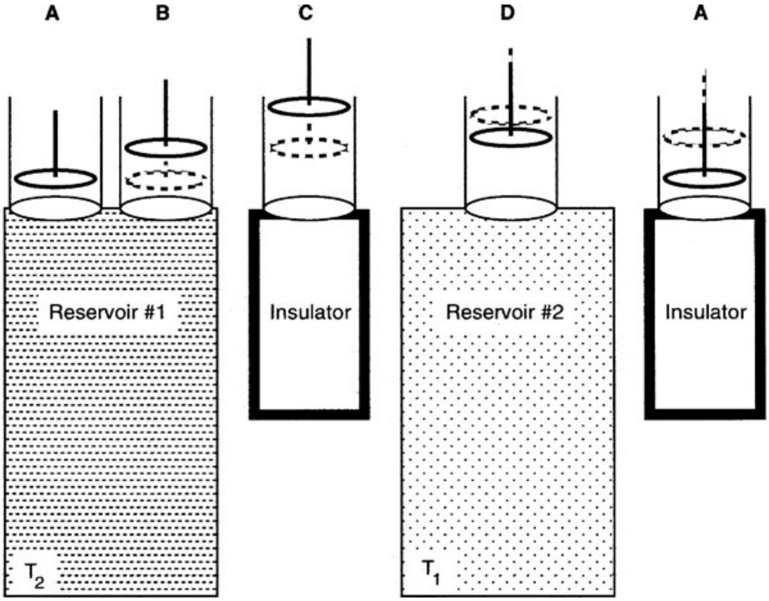


Fig. 11.6 The heat engine for the Carnot cycle thought experiment

the system fall. The system now resides at point *C,* which is on the isotherm of T_1. The piston is transferred into thermal contact with reservoir T_1, and a reversible isothermal compression is performed. As the system is compressed, heat flows from the system into the reservoir and a new state is established at point *D*. Finally, the piston is again thermally isolated and an adiabatic compression performed returning the system to T_2 and to point *A*.

Heat equivalent to q_2 is absorbed from reservoir T_2 during the isothermal expansion, and during the isothermal compression heat equivalent to $-q_1$ is absorbed from T_1. The net result of the two thermal transfers is $q_1 - q_2$. The work performed by the system (on the surroundings) is equal to the bounded area of the cycle, –*w*. We write $-w = q_2 - q_1$, which with rearrangement gives $q_2 = w - q_1$. Thus the heat delivered to the cold reservoir is the heat transferred from the hot reservoir minus the work performed. Only the part of the heat absorbed from the reservoir at the higher temperature is transformed in the Carnot cycle to work. The efficiency of a Carnot cycle heat engine is equivalent to the work performed divided by the heat extracted from the first reservoir:

$$\text{Efficiency} = \frac{-w}{q_2} \tag{11.16}$$

The remaining heat is not converted into work and ends up being discharged into the colder reservoir in step three of the cycle. From an efficiency standpoint, this heat is wasted. Therefore, the greater the heat drawn from the first reservoir compared to the heat placed in the second reservoir, the greater is the efficiency of the engine.

How can we relate the work available when heat is absorbed at T_1=hot and delivered at T_2=cold? With a little reflection we can appreciate this as a central question because this relationship in an ideal reversible engine will be independent of the materials from which the engine is built or that the engine acts upon. The available work is a universal function, an upper limit against which all real engines (or any system that uses energy) will be measured. To determine w we only need to know the values of $q_{ab} = q_2$ and $q_{cd} = q_1$ which can be found directly for our ideal gas system. For the isothermal expansion

$$\begin{aligned} q_{ab} &= -w_{ab} = -\left[-\int_{V_A}^{V_B} PdV\right] \\ &= nRT_{\text{hot}} \ln \frac{V_B}{V_A} \end{aligned} \tag{11.17}$$

and for the isothermal compression

$$q_{cd} = nRT_{\text{cold}} \ln \frac{V_D}{V_C} \tag{11.18}$$

Combining these equations we can write

$$\begin{aligned} -w =&q_{ab} + q_{cd} \\ =&nRT_{\text{hot}} \ln \frac{V_B}{V_A} + nRT_{\text{cold}} \ln \frac{V_D}{V_C} \end{aligned} \tag{11.19}$$

Algebraic rearrangement gives

$$\begin{aligned} \frac{q_{ab}}{T_{\text{hot}}} + \frac{q_{cd}}{T_{\text{cold}}} =&nR \ln \frac{V_B}{V_A} + nR \ln \frac{V_D}{V_C} \\ =&nR \ln \frac{V_B V_D}{V_A V_C} \end{aligned} \tag{11.20}$$

The far right-hand term needs to be solved if we are to have the relationship we are seeking. The linkage between the volumes of the cycle lies in recognizing that the work associated with the adiabatic expansion $V_B \rightarrow V_C$ and compression $V_D \rightarrow V_A$ can be found using, $dU = C_v dT$, and recognizing that the energy change is equal to the work, – PdV in an adiabatic process:

$$\begin{aligned} &\text{For} \quad V_B \rightarrow V_C \quad C_V \int_{T_{\text{hot}}}^{T_{\text{cold}}} \frac{1}{T} dT = C_V \ln \frac{T_{\text{cold}}}{T_{\text{hot}}} = nR \ln \frac{V_C}{V_B} \\ &\text{For} \quad V_D \rightarrow V_A \quad C_V \int_{T_{\text{cold}}}^{T_{\text{hot}}} \frac{1}{T} dT = -C_V \ln \frac{T_{\text{cold}}}{T_{\text{hot}}} = nR \ln \frac{V_D}{V_A} \end{aligned} \tag{11.21}$$

and

$$C_V \ln \frac{T_{\text{cold}}}{T_{\text{hot}}} - C_V \ln \frac{T_{\text{cold}}}{T_{\text{hot}}} = nR \ln \frac{V_B V_D}{V_A V_C} \tag{11.22}$$

or

$$nR \ln \frac{V_B V_D}{V_A V_C} = 0 \tag{11.23}$$

which is the relationship that we are looking for. We can now evaluate Eq. (11.20) and find that the expressions relating the heat transferred at the particular temperatures are equal to zero:

$$\frac{q_{ab}}{T_{\text{hot}}} + \frac{q_{cd}}{T_{\text{cold}}} = \frac{q_{\text{rev}}}{T} = 0 \tag{11.24}$$

The Carnot cycle goes through its process and returns to its original state with the resultant being zero. Since $\frac{q_{\text{rev}}}{T}$ fulfills the formal properties of a state function, it must be one. By dividing a path-dependent function by the absolute temperature, we have created a new state function, $\boldsymbol{S}$, or the *entropy*. Another way of looking at this is that when the path integral $\text{đ}q$ is integrated by an integration constant $\frac{1}{T}$ a state function is the result. Entropy has units of joules K^{-1}. The entropy for the reversible Carnot cycle can be written as

$$\Delta S = \frac{q_{ab}}{T_{\text{hot}}} + 0 + \frac{q_{cd}}{T_{\text{cold}}} + 0 \tag{11.25}$$

The efficiency of the machine will depend only on the temperature of the two reservoirs that we can connect from Eq. (11.16):

$$\frac{q_{ab}}{T_{\text{hot}}} = -\frac{q_{cd}}{T_{\text{cold}}}, \text{ therefore } \frac{q_{ab}}{q_{cd}} = -\frac{T_{\text{hot}}}{T_{\text{cold}}} \tag{11.26}$$

We can now find the efficiency by stirring these equations together:

$$\begin{aligned} \text{Efficiency} &= \frac{-w}{q_1} = \frac{q_{ab} + q_{cd}}{q_{ab}} = \frac{q_{ab}}{q_{ab}} + \frac{q_{cd}}{q_{ab}} \\ &=1 - \frac{T_{\text{cold}}}{T_{\text{hot}}} \end{aligned} \tag{11.27}$$

Consequently, for the efficiency of the inter-conversion of heat to work to be 100%, the temperature difference between the reservoirs must make the ratio $T_{\text{cold}}/T_{\text{hot}}$ zero. This occurs only as T_{cold} approaches zero (0 K) or as T_{hot} approaches infinity. This temperature scale, called the absolute or thermodynamic temperature scale, is obviously the same as the Kelvin temperature, which is derived from the pressure–temperature equation of state for an ideal gas. Because it is not possible in practice to achieve the conditions necessary to make this ratio zero, no machine operating in a cycle can achieve an efficiency of 100%, much less have an efficiency over 100%.

11.2.3 Entropy Is the Result of the Consideration of a Carnot Cycle

This is essentially the empirical proof of the second law of thermodynamics and we have watched the birth of a new state function, *entropy*, for our troubles. How do we get a good secure grasp on entropy in terms of what the state function means. Consider this train of thought. Entropy is a measure of the efficiency of our generalized ideal engine. The more efficient the heat engine the lower will be the entropy. We wish to perform a task with our machine, which means we want to go from point A to point B. Performing this job requires work that is simply task-directed energy. We choose heat as our energy source. The question is how well directed can we make the energy that appears as heat? We know that heat is randomly moving, scattered energy. If we trap each vectorial component directed toward the job at hand, we will extract all of the directed energy as work, but what of the vectors directed at odds to our task? The energy contained in these vectors is unavailable for work, will remain as random thermal energy, and will ultimately be discharged to the colder reservoir. The proportion of energy in heat available for work compared to the total energy is the efficiency.

Let us put this in more concrete terms. For simplicity, we consider our piston engine in which the movement of the piston can only move in one direction. The energy that enters the piston cylinder through the base is directed perpendicularly to the cylinder head, and so it is the process that moves the thermally energetic particle across the cylinder base that provides direction to the heat energy (i.e., the component that can do work). Any other component of energy that directs the particle to move from the perpendicular will not be able to appear as work but rather will remain as undirected energy or heat. Thus this driving force that causes the heat to flow from hotter to colder system is characterized by the entropy function. Since heat by definition has a certain undirected component, it is obviously impossible to convert all of the heat energy into work. Thus the efficiency of each cycle will never be 100%.

Now the meaning of the temperature difference should become clear. The consideration of temperature in the definition of entropy serves to normalize the amount of work–energy drawn from a heat cycle. The higher the temperature in the hotter reservoir the more likely will be the tendency to transfer heat and the higher the per degree ability of the transfer to produce work from the heat. Since the efficiency will be less than 100% even for a reversible engine and the wasted heat will be rejected into the cooler reservoir, there will be a driving tendency to move the heat into the cold reservoir. The colder this reservoir the greater the work per degree that can be drawn off to do work by the heat engine. For a reversible machine, the entropy driving the heat from hot reservoir into the heat engine will be equal in magnitude but opposite in sign to the entropy driving the waste heat into the colder reservoir.

The maximum energy is extracted by a reversible machine because the heat flow through the boundary is smooth in contrast to an irreversible machine in which there will be hot and cold spots within the boundary. These hot and cold spots will induce vectors of heat energy to become directed along the boundary thus scattering the already limited random heat energy to perform work. Irreversible machines and processes are characterized by this tendency to scatter energy that would otherwise be available in a reversible cycle, and thus on a per degree basis, more heat will be rejected as wasted, and less energy will be available to perform useful work. In the irreversible engine, the scattering of energy will decrease the efficiency of conversion of heat to work and increase the amount of heat delivered to the cold reservoir (higher entropy), thus increasing the overall entropy generated by the cycle and deposited in the Universe.

We generally can write for a reversible process

$$dS = \frac{đq_{\text{rev}}}{T} \tag{11.28}$$

For any other irreversible process

$$dS > \frac{đq_{\text{rev}}}{T} \tag{11.29}$$

The general case can be written after integrating

$$\Delta S = \frac{q_{\text{rev}}}{T} \tag{11.30}$$

Recalling the result for the work obtainable from a reversible isothermal expansion

$$w = -q = -nRT \ln \frac{V_2}{V_1} \tag{11.31}$$

rearranging gives

$$\frac{q_{\text{rev}}}{T} = nR \ln \frac{V_2}{V_1} \tag{11.32}$$

which is the same as writing

$$\Delta S = nR \ln \frac{V_2}{V_1} \tag{11.33}$$

for an ideal gas expansion. Consider the practical chemical and biochemical case of a system at constant pressure and recall that heat exchanged was equal to the enthalpy change of a system. Thus,

$$q_p = \Delta H \tag{11.34}$$

This enables equations of the following sort to be written

$$\frac{\Delta H_{\text{fusion}}}{T_{\text{melt}}} = \Delta S_{\text{fusion}} \tag{11.35}$$

This is the entropy of a phase change, where T is the melting point. Similarly, other equations can be written such as

$$\frac{\Delta H_{\text{vaporization}}}{T_{\text{boil}}} = \Delta S_{\text{vaporization}} \tag{11.36}$$

can be written. Calorimetry is a practical laboratory technique to experimentally determine these values for many materials of biological interest.

As a state function entropy is not conserved but rather generally increases in the Universe. At best, a reversible cycle will introduce no change in entropy for the system and surroundings, but any irreversible process will increase the entropy of the system plus the surroundings. Processes in which the combined entropy of the system and surroundings falls do not occur.

These general formulas will be of great value as shall be seen in the coming chapters. It is instructive and intellectually satisfying that these mathematical expressions also can be derived by a statistical approach. As pointed out in the introduction, it would be expected that the mechanistic approach of statistical mechanics would be able to provide the same mathematical formulation of thermodynamic law as does the empirically derived classical treatment.

We will now explore a mechanistic approach to these subjects. The statistical approach to deriving the state function of entropy, S, will lend mathematical form to the paraphrase that "a system will always move in the direction that maximizes choice." It will now be shown that the state function S can be defined in terms of a statistical variable, W

$$S = k \ln W \tag{11.37}$$

11.3 A Mechanical/Kinetic Approach to Entropy

We are familiar with the ideas underlying a microscopic treatment of a system of particles that leads to the equation of state for an ideal gas. In Chapter 7 we discussed the kinetic theory of gases and the equipartition theory. In Chapter 8 we saw how modifications of the classical limit are necessary, thus leading to quantum mechanics. What remains for us now is to codify these ideas in terms of state functions and to connect them in our minds to the macroscopic treatment of systems. As we know the treatment of a mechanistic theory is statistical in nature.

A central issue in physical studies and especially in biophysical chemistry is to connect macroscopic observables with knowledge of the structure and properties of the molecules composing the system. Much of the present day knowledge of molecular action in biological systems has been deduced from our understanding of the connections between macroscopic observables and microscopic properties at the level of fairly simple systems such as isolated gases and dilute solutions. Although we speak about the beauty of describing our system from a priori principles, the reality is that we have learned where to look at the molecular level because of the macroscopic behavior of a biological system. For this reason the study of ideal fluids and highly constrained systems is not only justified but necessary in the training of the biophysical scientist.

Keep in mind that equilibrium thermodynamics is concerned with the macroscopic properties of systems that are in equilibrium (i.e., the system has attained maximum entropy). Rate processes and transport phenomena (to be discussed in Section IV) are used to describe macroscopic systems moving toward, but not yet in equilibrium.

11.3.1 *The Statistical Basis of a Mechanistic Theory Is Reflected by System Properties*

We already have a good sense of which macroscopic empirical properties of state are linked to the mechanical properties of molecules. These we can summarize in Fig. 11.7. What is not obvious is how these properties are quantitatively connected. The arrow linking the properties represents the mapping functions that comprise statistical mechanics and thermodynamics. How do we approach the mapping processes that will allow us to derive the behavior of systems that depend on the behavior of groups of elements? Let us recast a fundamental question before

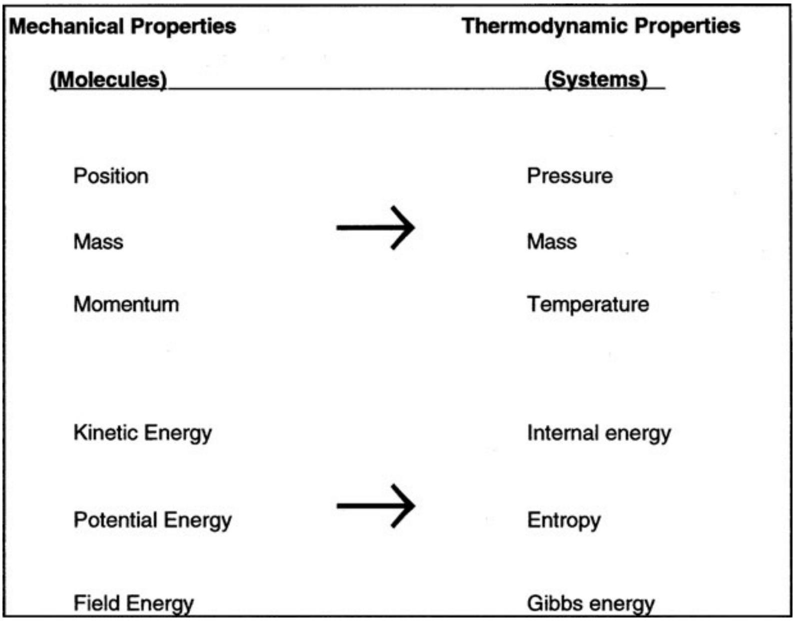


Fig. 11.7 Comparison between mechanical and thermodynamic properties

proceeding: Since systems are comprised of individual elements, are all of the properties of a system due to the additive properties of the elements, or is there some property of the system itself that is beyond the total of the individual parts? In other words, from the standpoint of "systemness," is the sum of the parts different from the whole they constitute? We have asked this question before, and the answers have always moved us toward a larger sense of what a system and a systemic interaction might imply for biological state spaces. Let us consider an example, namely, two ideal gases initially in separate containers. Because they are ideal they will not interact with one another at all. They are now connected by a tube and over time we will see that the particles move via diffusion into one another. The diffusion process is driven by the thermally induced random motion of the particles, and once the two gases are completely mixed they *never spontaneously unmix*. From a mechanical viewpoint there is no reason that any individual particle could not reverse its path and return to its "unmixed" position; yet as a group the particles in the system do not unmix. This spontaneous process of mixing irreversibly is the thermodynamic property of the system, entropy. Thermodynamic systems are inherently irreversible, and the driving force of this irreversibility, entropy, is a system property, not a property of an individual particle. Mechanical systems are inherently reversible because they are conservative, but entropy is not conserved. An interesting aspect of the mixing of the gas is that since the gas is ideal, the mixing has no relationship to the actual properties of the molecules. The mixing is driven entirely by entropy, and entropy is a property of the system (the group of molecules) alone.

If we restrict our thinking to quantum systems, we know that we cannot ascertain exactly the mechanical variables for all of the particles in a system. Furthermore, as we discussed earlier, even for a classical system in which it is theoretically possible to know in detail all of the mechanical properties of all particles, we can only realistically use a statistical approach that will provide us with a good sense of the possible and probable values for the particles taken as a whole.

11.3.2 Fluctuations Can Be Measured Statistically

We have alluded to the idea that macroscopic observables must be applied only when the system is large enough. The reason for this is mechanistic/statistical. We are not able to accurately describe too small a system. Following the example of Poincaré, consider the following system.

There are two connected containers that contain a number of particles. The probability that any single particle will be in either of the two containers at any instant is 1, while the probability that a single given particle is in a specific one of the containers is 1/2. The probability that each particle in the system, N, will be in one of the containers is 1/2 N. For a single particle, the chance that it would be in container A is 1/2 at any instant or 50% of the time. For two particles, the chance that both particles are found in A at one time is 1/4 or 25% of the time. For 10 particles, if container A is examined 1024 times, all 10 particles will be found in container A at the same time just once. However, if the number of particles in the system is increased to 100, the chance of finding all 100 particles in container *A* at one instant is less than 1 in 10^{30}. Even if the system is looked at again and again, each look taking a microsecond, it would take just over 4×10^{14} centuries to find all 100 particles in container *A* just once. For a fluctuation away from equilibrium (such as finding all the particles in container *A*), the larger the number *N*, the less likely it is that a statistically possible fluctuation will occur to invalidate the second law. For a molar quantity, large fluctuations simply do not occur in the real temporal framework of the Universe. For systems of large numbers of particles, the range of the fluctuation distribution will be quite narrow and the observable will be found to have a predictable and stable value.

11.4 Statistical Thermodynamics Yields the Same Conclusions as Classical Treatment of Thermodynamics

The goals of statistical thermodynamics are similar to those of classical thermodynamics in that we are interested in determining the internal energy of a system and gaining knowledge of the natural direction in which a system is moving over time. We describe a large system statistically because we cannot satisfactorily enumerate each particle, but we can describe the probabilities of a particular distribution of particles occurring. Furthermore, we often wish to know how these probabilities are changing with time; we need to know that the probability or likelihood of an event happening is well defined. We use the principle of *statistical convergence* to provide this definition. Statistical convergence occurs when a very large number of trials yield an increasingly predictable success rate. If for example the likelihood of an event is 17% for the first 500,000 trials, we can be quite confident that the next 500,000 trials will also have a 17% success rate.

11.4.1 The Ensemble Method Is a Thought Experiment Involving Many Probability Experiments

Accordingly, we use a statistical approach to codify internal energy and spontaneity because it is practically impossible to directly calculate the mechanical behavior of each particle in our system. The methodology used is the *ensemble method* devised by Gibbs. The ensemble is an imaginary set of probability experiments that employ the idea of statistical convergence to arrive at a probability or mean value associated with the ensemble. For example, consider the previously described, two-container system. Container *A* contains a single molecule and *B* is empty. We are told that at $t = 100$ ms, there is a 15% probability that the molecule will be found in *B*. What is the meaning of such a statement? In the ensemble method we imagine one of two experiments:

(1) We have set up a very large number of identical experiments and at $t = 0$, the stopcocks are all opened. At $t = 100$ ms, we look in container B (all of them). We find that in 15% of the experimental replicas, B contains the molecule.
(2) Alternatively, we can perform the same experiment on the same apparatus a very large number of times (the number is determined by the approach to statistical convergence) and we would we get the predicted result 15% of the time.

An ensemble is a Gedanken experiment with a large number, π, of imaginary systems each constructed to be a replica on a macroscopic or thermodynamic level of the real system whose properties we are investigating. If the thermodynamic state of a system containing 10^{20} molecules, N, can be characterized by the volume V and temperature T, the ensemble is constructed of π systems each duplicating the thermodynamic state (N, V, T). Although each system in the ensemble is thermodynamically identical, they are not identical on a molecular level. There are numerous different microscopic states that will give rise to the same thermodynamic state. This is not surprising because the abstraction of the thermodynamic state allowed us to use a very small number of observables to define the state, at the cost of having no detailed picture of the state itself. We would not expect three variables to be able to describe the details of 10^{20} particles.

Many different quantum states (stationary energy states for our purposes) will be represented at the specific instant in time when we construct our ensemble to replicate the thermodynamic state. For example, the pressure of our real system above, (N, V, and T), will be found by averaging the pressure of each imaginary system in the ensemble with equal weight assigned to each element of the ensemble. Thus an *ensemble average* is the equivalent to the thermodynamic variable.

There are two fundamental postulates upon which the superstructure of statistical thermodynamics is built. Considering our discussion to this point, they should be self-evident. Before reading further, try writing down what you think these might be. (Hint: we are attempting to build a model in which we wish to know if one abstract system is the same as another).

Postulate 1) The time average, over a long time period, of a mechanical variable M of a thermodynamic system is equal to the ensemble average as $\pi \rightarrow \infty$.

Postulate 2) An ensemble ($\pi \rightarrow \infty$) that represents an *isolated* thermodynamic system (N, V, E) will have systems distributed uniformly (i.e., with equal probability over all of the possible quantum states consistent with the specified values of N, V, E). This postulate is sometimes called “the principle of equal a priori probabilities.”

The first postulate depends on the second postulate since it is in the second postulate that the relative occurrence of the possible quantum states is defined. The first postulate also tells us that the ensemble average is an equilibrium condition, i.e., it is independent of time as $\pi \rightarrow \infty$.

Taken together these two postulates imply that the single isolated system of interest spends equal time in each of the available quantum states. This is called the *ergodic hypothesis* or assumption. The picture emerging is that the thermodynamic state of a system will be dominated by the microscopic occupancy of a series of quantum states that look alike. We have emphasized the equilibrium state, but invariably we are drawn to the query: what is happening on the way to the equilibrium state? You probably already have exposure to some aspects of the answer, which is one of dynamic or kinetic description. The equilibrium state microscopically is a dynamic steady-state equilibrium. Shortly we will take up this issue.

A close look at the second hypothesis reveals that this postulate applies only to isolated systems. Other systems exist and are of great importance. Three important systems are

(1) The isolated system where N, V, and E (energy) are given.
(2) The closed isothermal system of N, V, and T.
(3) The open, isothermal system defined by μ, V, and T (where $\mu =$ chemical potential).

Both N and μ represent sets of Ns and μs, making these systems applicable to the general case of more than one component.

The equivalent ensembles for each of these thermodynamic systems are respectively called; the *microcanonical*, *canonical*, and *grand canonical* ensembles. The method of derivation of these ensembles is to arrange the entire ensemble in such fashion that it appears to be isolated. Then both postulates may be used to determine the probability of finding a randomly selected system in a given quantum state. We will sketch the method briefly for a canonical ensemble without attempting a detailed derivation and then simply give the results for the microcanonical and grand canonical ensembles. The details of the derivations can be found in the texts listed in the end-papers. Other ensembles can be derived in addition to the three mentioned here that often have importance in biological systems. For example, ensembles in which surface area or the semi-permeable behavior of a boundary have been developed.

11.4.2 The Canonical Ensemble Is an Example of the Ensemble Method

The task is now to describe the process (ignoring many details) of the derivation of an ensemble. We will examine the canonical ensemble as if it were a research project with the goal to achieve maximum clarity.

SYSTEM: Fixed volume, fixed components – immersed in a very large bath of constant temperature.

GOAL: Calculate the mechanical variables such as energy and pressure.

METHOD: Determine the ensemble average of these variables (the first postulate).

> To do this we need to know (1) the value of a given variable in a given quantum state and (2) the fraction of systems in the ensemble.

CAVEAT: This system is not isolated. The energy levels of the system can fluctuate, and therefore quantum states belonging to different energy levels must be considered.

PROCESS: Since the prototypical system is immersed in a very large heat bath at T, each system of the ensemble must be likewise immersed. So

(1) The ensemble is built from π macroscopic systems of fixed V and N, each packed together in a lattice (see Fig. 11.8). The walls between each system allow only heat to pass (they are diathermal).

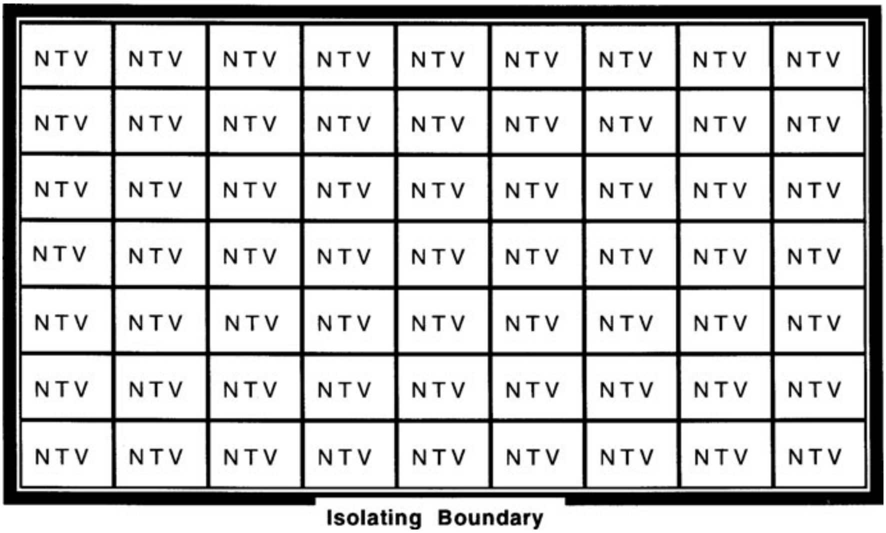
(2) Now place this entire lattice of systems in a giant heat bath of temperature T.

(3) Allow an equilibrium between the ensemble and the giant heat bath to be reached so that the ensemble is now at T.

(4) Insulate the outside of the entire ensemble (place an adiabatic wall around it) and remove it from the heat bath.

RESULT: > We now have an *isolated* ensemble at temperature T, with volume of πV, molecular number of πN, and total energy E_t.

> Each system in this ensemble is itself immersed in a heat bath of temperature T that is equal in size to $\pi - 1$. In other words the remaining systems in the ensemble act as the heat bath for any selected system we wish to examine.



NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV
NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV	NTV

Fig. 11.8 The canonical ensemble

THUS: The whole ensemble can be treated as an isolated "supersystem" and we can apply the second postulate to the canonical supersystem.

This quick sleight of hand seems a little unfair. But the second postulate tells us that every possible quantum state of this supersystem is equally possible and should be given equal weight in the calculation of the average values. Can we take advantage of this set-up in our consideration of the canonical system?

A single system in the canonical ensemble with N and V has allowed energy levels listed $E_1, E_2, \ldots, E_j$. Energy levels that are degenerate are listed multiple times. *Degeneracy* is defined as the number of distinguishable states of the same energy level. Each element in the supersystem has the same list of eigenvalues though the distributions over the E_j will be different. Since every element of the supersystem is fixed with respect to its N, V, and T, the canonical ensemble is really just a degenerate microcanonical ensemble. We are interested in delineating the number of elements, n_j, in a specific energy state, E_j for a supersystem with total energy E_t. This is our goal because we know that the thermodynamic state is going to look like the set of energy states that are most heavily occupied. In other words the most probable distribution and those distributions that are indistinguishable from it are going to dominate the set of distributions and hence the observed state of the supersystem (i.e., the state of the canonical ensemble).

The total energy of the ensemble is

$$E_t = \sum n_j E_t \tag{11.38}$$

and the number of elements is

$$N = \sum n_j \tag{11.39}$$

We wish to look at a given moment and find the number of systems (elements) in each energy state, i.e., n_1 in E_1; n_2 in $E_2 \ldots n_j$ in E_j. The set of numbers $n_1, n_2, \ldots n_j$, is a distribution set, $\mathbf{n}$. There are a large number of ways that the systems of the ensemble can be assigned to E_j. This number of ways, W, is determined by Eq. (4.26), which is used for a distribution of distinguishable systems.

$$W(n) = \frac{N!}{\prod_{j=1}^{r} N_j!} \tag{11.40}$$

The elements of each ensemble are macroscopic, and we could label each of them should we wish.

Our goal of determining thermodynamic variables from a mechanical perspective is in reach. We now select a system at random from the canonical ensemble. What is the probability p_j that it will be in state E_j? The answer is that p is the average value of n_j divided by the total number of systems, N. The value of $\overline{n_j}$ depends upon which of a wide variety of plausible distributions is chosen. There are many distributions that will generally fulfill the constraints of Eqs. (11.38) and (11.39). For any specific distribution, the fraction of elements in E_j is n_j/N. The overall probability will be found by averaging each n_j/N over all the allowed distributions, with each weighted equally according to the principle of equal a priori probabilities. Thus we can write

$$p_j = \frac{\overline{n_j}}{N} = \frac{1}{N} \frac{\sum\limits_n W(\mathbf{n})\, n_j(\mathbf{n})}{\sum\limits_n W(\mathbf{n})} \tag{11.41}$$

We now have the tools to calculate the canonical ensemble average for any general mechanical property of a supersystem:

$$M = \sum_j M_j p_j \tag{11.42}$$

These last two expressions are the machinery that the arrow in Fig. 11.7 represented. They are practically difficult because the summations are mathematically formidable. However if we let $\pi \rightarrow \infty$ we can employ the principles of continuous probability outlined in Chapter 5. If the variables are made large, we know that multi-nomial coefficients such as $W(\mathbf{n})$ will become very sharply peaked at their maximum. We can allow the n_j's to become arbitrarily large because N is approaching infinity. In the limit we can legitimately consider only the most probable distribution, $\mathbf{n}^*$, and thus only the weight $W(\mathbf{n}^*)$ is included. We now write

$$p_j = \frac{\overline{n}_j}{N} = \frac{n_j^*}{N} \tag{11.43}$$

This problem as well as its mathematical expression is similar to the problem of the Boltzmann distribution which is used to determine the most probable distribution of a large number of particles among the energy levels available to them. Thus

$$\frac{n_j^*}{N} = \frac{g_j e^{-E_j/kT}}{Z} \tag{11.44}$$

where we define Z to be the *canonical partition function*:

$$Z(V, T, N) = \sum_j \frac{g_j e^{-E_j/kT}}{Z} \tag{11.45}$$

We have introduced a weighting factor g_j which is the statistical weight of the energy level E_j. This factor is a measure of the degeneracy of a particular energy level.

The links between the quantum mechanical energy states of a macroscopic system and the thermodynamic properties of the system are made through the partition functions. For the canonical ensemble the average energy $\overline{E}$ can be written in terms of Z and equated with the thermodynamic internal energy U:

$$\overline{E} = kT^2 \left(\frac{\partial \ln Z}{\partial T} \right) \tag{11.46}$$

This discussion is not a detailed derivation of statistical mechanics or its applications but has been structured so that the reader may have a sense of how the links between a mechanistic and thermodynamic theory are forged. A variety of systems can be treated with statistical techniques provided a partition function can be found. Unfortunately, this is often difficult with the complex molecules and with the open systems of biological relevance. However, the application of these statistical methods is the basis for many important qualitative and semi-quantitative treatments of systems of biological interest.

11.4.3 The Distribution of Energy Among Energy States Is an Important Description of a System

We have introduced the idea of the energy distributions in the systems. The similarities to the energy states of a molecule are important. As we discussed in some depth in Chapter 8 the quantum states of a system are not continuous but are discrete. Let us take a moment to explore the arrangement of energy states in quantum systems. Quantum systems do not contain internal energy of arbitrary value but rather have a well-defined set of energy states. A one-dimensional harmonic oscillator has energy states given by $E_n = \left(\left(n + {}^1\!/_2 \right) hv \right)$, where n is an integer greater than or equal to 0. The ground-state energy is assigned the zero-point energy, and energy is added in fixed quanta of hv. These energy levels will be equally spaced. The energy levels of the hydrogen atom are more complicated and are described by the quantum numbers n, l, and m as well as spin numbers. The spacing of these energy levels is not equal but is instead $\approx \frac{1}{n^2}$. In a system like the hydrogen atom we begin to see degeneracy because certain distinct energy levels have the same energy. When we now consider the energy states of a complete thermodynamic system such as a liter of a liquid, the practical problems become much larger. Though the fundamentals remain intact and a very large number of definite energy levels can in principle be calculated, this is impossible in practice. For such large systems we use approximate energy levels for two reasons. First, even if the system starts in definite energy states (the position and momenta of a group of gas molecules), over a period of time collisions between the particles will lead to exchange of momentum and the states of definite momentum will be altered. Thus the collisions are *perturbations* or forces that are ignored in the initial abstraction and that lead to scattering of the energy states. The second reason for using approximate levels is that for large systems the energy states are very closely spaced. So small are the separations that it becomes impossible to speak precisely of separate defined states. While the spacing for a small oscillator is on the order of 10^{-20} J, the spacing for a large system will be $10^{-10^{23}}$ J. It becomes more sensible to treat the system as if it were a continuous distribution, and therefore we speak of the *density of states in energy*, $g(E)$, or $^{dn}/_{dE}$. For a large system, the density of states is a smooth function of the energy on the atomic scale.

What is the relationship between the density of states and energy in a large system? The answer can be deduced by considering a greatly simplified system consisting of 50 equivalent oscillators. This is a problem of determining the number of ways we can distribute quanta between energy levels (all are indistinguishable). The general formula is

$$W = \frac{(Q+N-1)!}{Q!\,(N-1)!} \tag{11.47}$$

If we add no quanta of energy to the system all the oscillators are in their ground state, and there is only one possible arrangement. If a single quanta is added there are 50 possible states. Two quanta give 1275 possible states, and so on as enumerated in Table 11.1. The addition of quanta leads to a very rapid rise in the density of states (which is $^{W}/_{h\nu}$) since g is multiplied by the factor $\frac{(Q+N)}{(Q+1)}$. *The density of states rises in an exponential fashion with energy*.

Table 11.1 The number of arrangements or ways that Q quanta can be arranged in 50 identical indistinguishable oscillators

Quanta (Q)	Arrangements (W)
0	1
1	50
2	1275
3	22100
4	292825
5	3162510
10	6.28×10^{10}
20	1.16×10^{17}

11.4.4 Heat Flow Can Be Described Statistically

How do we understand heat flow in statistical terms? We proceed using the following model from Waldram (1985). There are two systems A and B free to exchange heat but with no other significant interactions; thus their individual energy states remain unaltered by the contact. We can graphically describe this joint system (Fig. 11.9) by enumerating the quantum states of system A on the x axis (E_m) and those of B on the y axis (E_n). System A contains 5000 identical oscillators and system B contains 10,000 oscillators. The total energy of the system is 15,000 quanta. The joint system has states in which both A and B are in definite states, m and n, and the joint state has energy E_{mn}. On our graph the joint states are represented by dots. The total density of joint states will be equal to the product of the density of states for A and B, $g_A g_B$. The interaction between the systems occurs over a limited range of energy δE to which both A and B have accessibility. (They will be able to exchange energy in this limited range governed by the rules of quantum kinetics.)

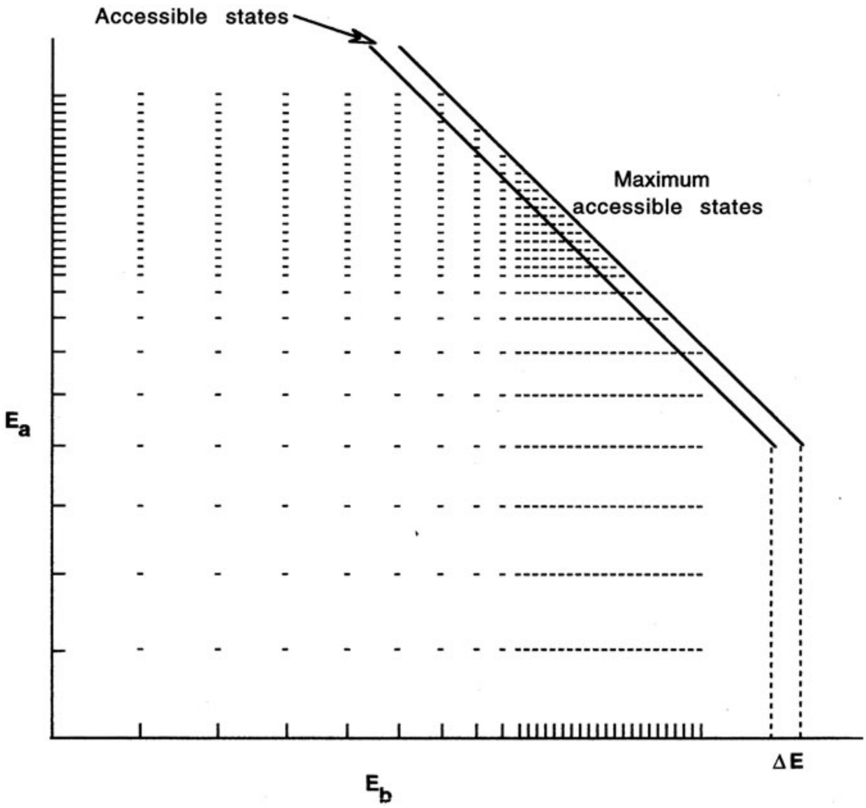


Fig. 11.9 The arrangement of the quantum states of a joint system to investigate heat flow (Modified from *The Theory of Thermodynamics* by J.R. Waldram. Copyright 1985 by the Cambridge University Press. Reprinted by permission of Cambridge University Press.)

The dots lying between the sloping lines are the accessible states for the joint system. The density of joint states is clearly not uniform in this band of interaction, and ultimately the region with the highest density of states would be expected to dominate. Therefore we would expect heat to flow so that the value of $g_A g_B$ increases. This can be seen in Fig. 11.10. We begin our joint system at point *a*. The value of $g_A g_B$ in the direction of *b* is 10^{100} larger than at *a* and 10^{200} larger than at b'. Any significant movement toward b' will be negligible when compared to the number of states lying in the direction of *b*. The number of jumps into the states of highest density (i.e., toward *b*) is much more likely to be seen than the reverse set of jumps out of these more abundant states. Early in the process only several jumps may have occurred and then the possibility of a jump toward b' and away from *b*, thus indicating the movement of heat in the "wrong" direction will be quite high. However,

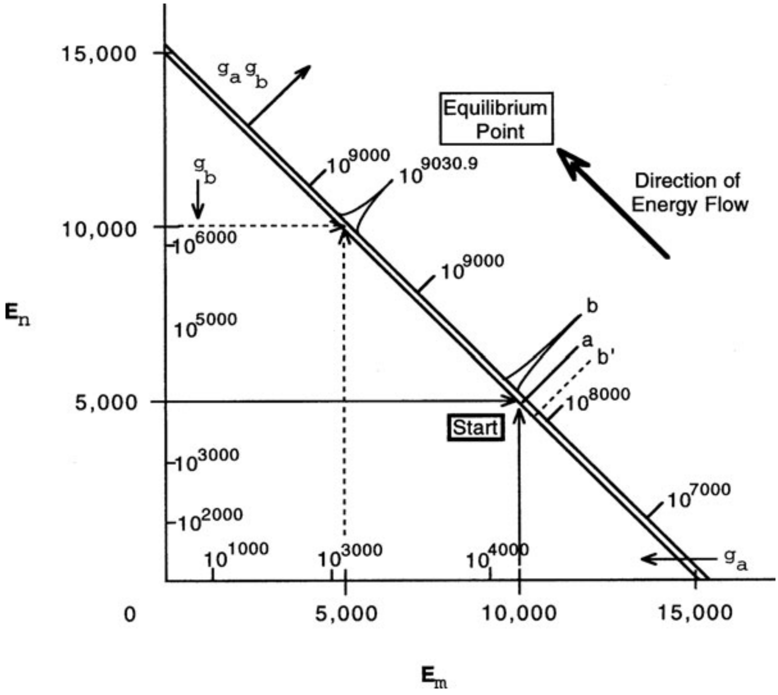


Fig. 11.10 Heat flow as a consequence of the interaction of the two systems. Point *c* referred to in the text is labeled "Equilibrium point" in the figure (Modified from *The Theory of Thermodynamics* by J.R. Waldram. Copyright 1985 by the Cambridge University Press. Reprinted by permission of Cambridge University Press.)

by the time 1500 jumps have occurred, 1000 will have occurred in the direction of b and only 500 in the direction of b'. At this point system A will have the value of 9500 ± 36 quanta with a narrow distribution indicating little uncertainty in the amount of heat flow from A to B. The probability that heat flows the "wrong" way for this system is about 10^{-38} represented by the tail on the wrong side of b. The probability of a significant heat flow to b' is about 10^{-150} for this system. Thus it is extremely unlikely that such an event would ever be seen in the lifetime of the Universe. The flow of heat from a hot to a cold system is a statistical phenomenon. By the time 100,000 jumps have occurred the joint system will have moved to point c on the graph. At c each oscillator in the joint system will have 1 quanta, and so system A will have E_A equal to 5000 ± 82, with the standard deviation representing the thermal fluctuations around the equilibrium point. Therefore, the flow of heat from a colder to a hotter body is *not* prohibited as the Clausius statement of the second law of thermodynamics states. It is just that such an event is so unlikely that we can safely ignore it as ever having a probable occurrence.

11.4.5 Internal Molecular Motions, Energy and Statistical Mechanics Are Related by a Partition Function

It would be useful to know the partitioning of the internal energy of a molecule among its various energy modes, i.e., translation, rotation, and vibration. A partition function for internal molecular motions will relate the thermodynamic properties to the contribution of the internal degrees of freedom:

$$\mathrm{Z}_t = \sum g_j e^{-E_j/kT} \tag{11.48}$$

It is a reasonable approximation to take the internal energy as a sum of each of these independent terms:

$$E_I = E_t + E_r + E_v \tag{11.49}$$

Using theoretical formulas derived via quantum mechanical energetics we can evaluate the different contributions to the molecular partition function:

$$Z_I = Z_t\, Z_r\, Z_v \tag{11.50}$$

We will not derive the expressions for the molecular partition functions in this text but the formulas are listed in Table 11.2. The interested reader can find derivations in the texts listed in the supplemental reading section.

Table 11.2 Molecular partition functions

Type of motion	Degrees of freedom	Function
Vibrational	1	$\dfrac{1}{1-e^{-h\nu/kT}}$
Rotational (linear molecule)	2	$\dfrac{8\pi^2 IkT}{\sigma h^2}$
Rotational (non-linear molecule)	3	$\dfrac{8\pi^2(8\pi^3 ABC)^{1/2}(kT)^{3/2}}{\sigma h^3}$
Translational	3	$\dfrac{(2\pi mkT)^{3/2}V}{h^3}$

The terms have their usual meanings and *s* is a symmetry number that is equal to the number of indistinguishable positions into which a molecule can be turned by rigid rotation, A, B, C, and I are moments of inertia

11.5 Entropy Can Be Described and Understood on a Statistical Basis

We have noted that the state of equilibrium for a macroscopic system is reached when the state functions no longer change with time. The equilibrium position is the most probable arrangement of the particles that make up the system. For a large system there are a very large number of degeneracies. Thus, finding the equilibrium position is the same as finding the most probable distribution of the particles in a system. In this view of entropy, it is necessary to be able to express the possible distributions of a set of particles among states of a defined energy level. The number of particles is given by N_1, the energy level may be denoted as E_1, and the degeneracy of this energy level is defined as g_1. The question with regard to distributions is, How many possible ways can N_1 particles be distributed into g_1 boxes at energy level E_1? Depending on whether or not certain restrictions apply to the particles, that is, (a) whether there are characteristics that distinguish the particles comprising N_1 or (b) whether there are rules that specify how many particles may be placed in each box, certain mathematical expressions may be written to describe the distributions.

11.5.1 Different Statistical Distributions Are Needed for Different Conditions

Where the particles are distinguishable but no restriction applies to the number of particles found in each box, the distribution is defined by Boltzmann statistics:

$$t_1 = g_1^{N_1} \tag{11.51}$$

where t_1 is the number of ways that the N_1 particles can be arranged in g_1 boxes. In describing a complete system, there are N_1 particles distributed in g_1 boxes, N_2 particles distributed in g_2 boxes, N_3 particles distributed in g_3 boxes, and so on, where the different subscripts are used to designate different energy levels. The population of particles in this system will be given by N, where

$$N = N_1 + N_2 + N_3 + \mathrm{K} = \sum_i N_i \tag{11.52}$$

The total number of ways, T, that the entire population, N, can be distributed will be given by

$$T_{\text{Boltzmann}} = N! \prod_i \frac{g_i^{N_1}}{N_i!} \tag{11.53}$$

This equation does not descend directly from the previous notation. The conditions for Boltzmann statistics are based on the distinguishable nature of each particle. Because of this condition, many arrangements can be made by forming subgroups from the population *N*. These subgroups must be taken into account in overall summation. In general, the number of ways to group *N* objects into subgroups n_1, n_2, $n_{3,}\ldots$ is given by

$$\frac{N!}{n_1!n_2!n_3!n_i!} \tag{11.54}$$

Boltzmann statistics are based on assumptions that are not consistent with the generally regarded dogmas of quantum mechanics. Other statistical distributions are possible. When the particles comprising *N* are considered to be indistinguishable but no limit is placed on the number of particles contained in each box, *Bose–Einstein statistics* result. *Fermi–Dirac statistics* result when the particles are indistinguishable and there is a limit of one particle per box. Both of these conditions are actually more accurate in the world of real particles than Boltzmann statistics. Fermi–Dirac statistics apply to particles that obey the Pauli exclusion principle, such as electrons and positrons, whereas Bose–Einstein statistics apply to particles that have *integral spin*, such as photons, and are not subject to the Pauli exclusion principle. Boltzmann distributions in fact do not relate to any real particle. It can be shown, however, and it is left as an exercise for the reader to do so, that for a sufficiently dilute system, where $N_i << g_i$, the *T* for Fermi–Dirac and Bose–Einstein statistics are equal to $T_{\text{Boltzmann}}/N!$. This results in the following expression:

$$T_{\text{special Boltzmann}} = \frac{T_{\text{Boltzmann}}}{N!} = \prod_i \frac{g_i^{N_1}}{N_i!} \tag{11.55}$$

The advantage of this approach is that the mathematics of this modified expression is easier to work with than those of the Fermi–Dirac and Bose–Einstein statistics.

To review, the system for which a most probable distribution of states is sought is generally constrained in several ways. First of all, it is an isolated system, and so its energy, *U*, is constant. Furthermore, in an isolated system at equilibrium, the total number of particles, *N*, may also be considered constant. The result is

$$\frac{N_i}{N} = \frac{g_i e^{-U_i/kT}}{Z} \tag{11.56}$$

This is the *Boltzmann distribution*, where *k* is Boltzmann's constant and *Z* is the partition function, a function that sums over all the system's energy levels. *Z* is defined as follows:

$$Z = \sum_i g_i e^{-U_i/kT} \tag{11.57}$$

Often, what is of most interest is the ratio of the number of particles or molecules in two different energy levels, and this is given by

$$\frac{N_i}{N_j} = \frac{g_i}{g_j} e^{-(U_i - U_j)\big/kT} \tag{11.58}$$

This formulation indicates that lower energy states will generally have more molecules than higher ones, assuming that a rapid increase in degeneracy with increased energy does not occur.

11.5.2 Phenomenological Entropy Can Be Linked to Statistical Entropy

Earlier we calculated *W* and found that *W* was maximal at equilibrium. If *W* is maximal at equilibrium, then it follows that for a system moving spontaneously in a direction toward equilibrium, the change in *W* from state 1 to state 2 will be given by

$$W_2 - W_1 \geq 0 \tag{11.59}$$

In this statement, *W* has the form of a state function. It is an extensive state function and is one that increases in a spontaneous process. Both of these are qualities for the state function of entropy, *S*.

The state functions *S* and *W* can be related

$$\Delta S = k \ln W \tag{11.60}$$

and

$$\Delta S = k \ln \frac{W_2}{W_1} \tag{11.61}$$

where *k* is Boltzmann’s constant.

It can be shown that the statistical expression of entropy, S_{se}

$$S_{\text{se}} = k \ln W \tag{11.62}$$

is equal to the traditionally defined entropy

$$S = \frac{q_{\text{rev}}}{T} \tag{11.63}$$

and if the expansion of an ideal gas is considered, it can be shown that

$$\Delta S = k \ln \frac{W_2}{W_1} = Nk \ln \frac{V_2}{V_1} \tag{11.64}$$

11.6 The Third Law of Thermodynamics Defines an Absolute Measure of Entropy

The *third law* of thermodynamics states that at absolute zero the entropy of all pure, perfect crystals is 0:

$$S_{\text{pure-perfect xtal}}(0\,\text{K}) = 0 \tag{11.65}$$

This makes some sense, if it is considered that it is possible to imagine that a crystal at zero degrees would be perfectly arranged so that perfect order reigned everywhere, and hence the location of every atom could be known. Such a system would be like that of the three particles when they were all in energy level 0. Generally, only a crystal could have this structure, not a liquid or a gas. Also, if the system were not pure, the entropy would not be at a minimum because by separating the components of the mixture, the entropy would fall. (The entropy of mixing will be taken up shortly.)

The temperature dependence of entropy is easily found:

$$\Delta S = S_2 - S_1 = \int_{T_1}^{T_2} \frac{q_{\text{rev}}}{T} dT \tag{11.66}$$

At constant pressure, $q = H$ and $\Delta H/T = S$, while $\Delta H = C_p \Delta T$, so that

$$\Delta S = \int_{T_1}^{T_2} C_{\text{p}} dT \tag{11.67}$$

this is equal to

$$\Delta S = C_{\text{p}} \ln \frac{T_2}{T_1} \tag{11.68}$$

Since C_{p} is always positive, the entropy always increases as the temperature is increased. This is a result that makes sense when considered in terms of the three-particle example, where the addition of energy led to a new equilibrium with a larger W. The tendency of systems to move in the direction of greatest choice as characterized by entropy is important throughout the biological Universe.

Further Reading

Statistical Thermodynamics

Dill K.A. and Bromberg S. (2003) *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology.* Garland Science, New York.

Hill T.L. (1986) *An Introduction to Statistical Thermodynamics.* Dover Publications, New York.

McQuarrie D.A. (1973) *Statistical Thermodynamics.* University Science Books, Mill Valley, CA.

For a derivation of the molecular partition functions, see Moore W.J. (1972) *Physical Chemistry*, 4th edition. Prentice-Hall, New York, Chapter 14.

Articles of Interest

Bennett C.H. (1987) Demons, engines and the second law, *Sci. Am.*, **257, 5**:108–112.

Mulero F.C. and Rubio P. (1994) The perturbative theories of fluids as a modern version of van der Waals theory, *J. Chem. Educ.*, **71:**956–962. (This article relates the ideas of intermolecular forces, gases, liquid properties and the statistical mechanical tool of perturbation theory. A nice historical context is given to the development of the ideas.)

Problem Sets

1. Consider a three compartment container into which the same particles described in Section 11.3.2 are placed. Calculate the probability of 1 particle being found in the middle compartment, 10 particles, 100 particles. They re-sort themselves every pico-second. Assuming that you have a machine to look at the compartments every 10^{-12} s, how long would you need to observe the system to find all of the particles in the middle container? What if they sorted every femtosecond?
2. Patch-clamping techniques have shown that individual ion gates in the resting membrane open and close in a random fashion. Propose a method by which the observation of a single ion gate may be used to predict the state of the resting or steady-state membrane, that is, gates open or closed. Propose a second method to find the resting state of the membrane.
3. An inventor sends a patent application to the Patent office. In it (s)he claims to have found a chemical compound ensures eternal youth for all who drink it. The patent office rejects it on the basis that the formulation is a perpetual motion machine. Explain why. Did they reject it because it was a perpetual motion machine of the first or second kind?
4. An inventor comes to you with a proposal for a new machine that appears to be a perpetual motion machine. Before you throw him out, he begs your indulgence for a moment. Against your better judgment, you relent. He proposes to use the positive volume change of ice on freezing to power a machine. His machine works as follows. A quantity of liquid water is located in a cylinder that has a piston and a cooling unit associated with it. Initially, the cooling unit freezes the water, and as the ice expands, work is done on the piston by raising it. This mechanical work is converted to electrical power by a generator turned by the piston's movement and stored in a battery. As the ice melts, the piston moves downward, the movement again being captured by a differential gear and turning the generator. At the instant the ice becomes liquid, i.e., 273 K, the cooling unit comes on, powered by the battery. As the ice again forms, the piston moves and work is done, and so on. Should you invest in this invention? Why or why not?

5. Perfect crystals at absolute zero are defined by the third law of thermodynamics to have an entropy of 0.

   a. What does this imply about imperfect crystals at absolute zero?
   b. What are the molecular events that lead to a rise in entropy as the temperature of the perfect crystal increases?
   c. If a typical protein is crystallized and then cooled to absolute zero, do you think that it will have an entropy of zero?

Chapter 12
Which Way Is That System Going? The Gibbs Free Energy

Contents

12.1 The Gibbs Free Energy Is a State Function that Indicates the Direction and Position of a System's Equilibrium

So far we have treated the first and second laws separately, but in a real system both energy conservation and entropy play a role. The role of entropy in determining the spontaneous direction of a system was derived for an isolated system that exchanged neither energy nor material with its surroundings. There are few biological systems of interest that fulfill the requirements of this isolation. We need a state variable

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_12, © Springer Science+Business Media, LLC 2010

that will indicate the direction and equilibrium position of a system that undergoes radical changes in energy (usually measured as enthalpy) and entropy together. This state function exists. It is the *Gibbs free energy*, ***G***.

The Gibbs free energy is applicable to systems that are constrained to be at constant temperature and pressure and therefore is ideally suited to biological systems. G is an extensive variable that relates enthalpy, entropy, and temperature in such a fashion that the spontaneity of a system may be determined:

$$G = H - TS \tag{12.1}$$

If ΔG for a change in a system is negative, then that change can occur spontaneously. Conversely, a change in state with a positive ΔG is associated with a non-spontaneous process. It follows therefore that when ΔG has a value of 0, the system is at equilibrium. Like enthalpy and energy, G has units of joules in the SI system. It is necessary to stress that ΔG depends only on the initial and final states of the system under study. It does not provide information about the rate or the kinetics of the change.

The point in having a variable such as G is to determine the "natural" direction or spontaneous direction of a system undergoing a change toward equilibrium. Differentiating Eq. (12.1) gives

$$dG = dH - T\,dS - S\,dT \tag{12.2}$$

The restrictions of constant temperature ($dT = 0$) and pressure (which gives $dH = \text{đ}q$) are now applied:

$$dG = \text{đ}q - T\,dS \tag{12.3}$$

Entropy can be defined as $dS = \dfrac{\text{đ}q_{\text{rev}}}{T}$, and this gives the interesting result:

$$dG = \text{đ}q - \text{đ}q_{\text{rev}} \tag{12.4}$$

Now if

$$\text{đ}q = \text{đ}q_{\text{rev}} \tag{12.5}$$

then $dG = 0$ (at constant temperature and pressure).

In the last chapter, a reversible process was defined as one that is at equilibrium, so this solution fulfills the criterion that equilibrium is defined when ΔG is 0. For the other case, in an irreversible process

$$\text{đ}q < \text{đ}q_{\text{rev}} \quad \text{so } dG < 0 \tag{12.6}$$

(See Appendix K for the derivation of $\text{đ}q = \text{đ}q_{\text{rev}}$.)

These results also make sense since any real or natural process has already been demonstrated to be an irreversible one, and a negative value for ΔG is expected for such a case.

The relationships derived for the Gibbs free energy are general and applicable for any thermodynamic system appropriately constrained by temperature and pressure. The analysis has advanced enough at this point so that it is desirable to recall that all of the derivations so far have been for “ideal” systems. Very shortly, the real and “non-ideal” world will necessarily be confronted, and equations will need to be derived for that eventuality. It will turn out that the forms of the simple equations derived so far can be kept even though the forces operating in non-ideal systems are more complex than those in the ideal systems.

Thermodynamic systems have many functions that can define the state. In the system considered so far, the state has been adequately defined by the pressure (P), temperature (T), and volume (V), which have allowed the determination of the fundamental properties of energy (U) and entropy (S) as defined by the laws of thermodynamics. By taking into account various constraints, two useful, derived state variables have been defined, namely enthalpy (H) and Gibbs free energy (G). Other derived variables exist, such as A, the Helmholtz free energy, but will not be considered here. For the most part, state functions presented will be adequate for the remaining discussion.

12.2 The Gibbs Free Energy Has Specific Properties

In the formulation of the Gibbs free energy, both temperature and pressure are constrained to be constant. While such a simplification is useful in many cases, there are also many situations where it is valuable to know what the free energy of a system will be when changes of either temperature or pressure occur. A very practical reason to know such a relationship lies in the simple fact that standard tables of free energy (as well as enthalpy and entropy) are written in the standard state. Clearly, if a scientist is interested in a system at 310 K and has tables that provide free energies in a standard state at 298 K, then to be able to use the tabulated values, the temperature dependence of G needs to be considered and calculated. Beyond this practical (and time-saving) reasoning, there are many questions in biological systems that need to consider altered pressures and temperatures. For example, what effect does the increased temperature of a fever have on the spontaneity and equilibrium of the steps in a metabolic process? How does an increase in temperature affect the free energy of a conformational change in a protein or a lipid bilayer? How about the effect on metabolic processes when the temperature drops as in a hibernating bear or in a cold water immersion (children have, for example, been “drowned” for greater parts of an hour in icy water, appeared clinically dead, and been brought back to life with apparently no serious side effects)? The role of pressure in a biological system comes into play when considering gas exchange in the lung and tissues, both events that do not occur at standard or constant pressures. The *oncotic* pressure of the body fluids is crucial for the normal retention of intravascular water in the high hydrodynamic pressure environment of the bloodstream; loss of protein in protein-wasting nephropathies can lead to drastic pathologies. What, then, are the properties of *G*, related to the variables of state so far considered?

The problem can be approached by starting with the definition of *G*:

$$G = H - TS \tag{12.7}$$

This is then differentiated, which gives the following, the fundamental equation:

$$dG = dH - T\,dS - S\,dT \tag{12.8}$$

A number of substitutions can be made. First, because $dH = dU + P\,dV + V\,dP$, the following is written:

$$dG = dU + P\,dV + V\,dP - Td\,S - S\,dT \tag{12.9}$$

By the definition of the first law, dU can be written $\text{đ}q + \text{đ}w$ so

$$dG = \text{đ}q + \text{đ}w = P\,dV + V\,dP - T\,dS - S\,dT \tag{12.10}$$

For a reversible process, it can be shown that $\text{đ}q = TdS$, and it can be assumed that $\text{đ}w$ is the *PV* work or $-P\ dV$, if only pressure–volume work is performed. Substituting then gives

$$dG = T\,dS - P\,dV + P\,dV + V\,dP - T\,dS - S\,dT \tag{12.11}$$

Combining the terms gives the following result:

$$dG = V\,dP - S\,dT \tag{12.12}$$

This formulation provides the dependence of *G* on temperature and pressure.

Equation (12.12) is appropriate when only pressure–volume work is being considered. As previously stated, much of the work done in or by a cell or organism takes forms other than *PV* work. Therefore, a new term, w_{other}, can be added and defined as needed in various systems:

$$dG = V\,dP - S\,dT + dw_{\text{other}} \tag{12.13}$$

Again, considering Eq. (12.12), an equivalent equation can be written using partial derivatives:

$$dG = \left(\frac{\partial G}{\partial P}\right)_{\mathrm{T}} dP + \left(\frac{\partial G}{\partial T}\right)_{\mathrm{P}} dT \tag{12.14}$$

This equation also relates the change in free energy to changes in pressure at constant temperature and to changes in temperature at constant pressure. An important result of having both Eqs. (12.11.) and (12.14) is that the following identities are revealed:

$$\left(\frac{\partial G}{\partial P}\right)_{\mathrm{T}} = V \tag{12.15}$$

$$\left(\frac{\partial G}{\partial T}\right)_{\mathrm{P}} = -S \tag{12.16}$$

We emphasize again that the way to make effective use of partial differentials is to identify their physical meaning. In the case of Eq. (12.15), volume is always positive, so the free energy of a system at constant temperature always increases as the pressure increases. On the other hand, Eq. (12.16) indicates that the free energy of a system decreases as temperature increases in a system at constant pressure. This result occurs because entropy is always positive. Equation (12.16) provides the answer to the problem of finding the temperature dependence of free energy. A number of further manipulations of Eq. (12.16) are possible, the use of each depending on whether or not it is useful for a particular problem. One of the most useful expressions is the *Gibbs–Helmholtz equation*, which is used in situations where enthalpy or entropy varies with temperature:

$$\left(\frac{\partial\left({}^{G}\!/_{T}\right)}{\partial T}\right)_{\mathrm{P}} = -\frac{H}{T^2} \tag{12.17}$$

For the effect on G of a change from T_1 to T_2, the following version of Eq. (12.17) is of value:

$$\frac{\Delta G_{(T_2)}}{T_2} - \frac{\Delta G_{(T_1)}}{T_1} = \int_{T_1}^{T_2} \frac{\Delta H_{(T)}}{T^2} dT \tag{12.18}$$

In many biological situations, the temperature dependence of entropy or enthalpy, and hence of the free energy, is minimal. In such cases, usually when the temperature difference between states is not very great, the difference can be ignored, and the following approximation can be used:

$$\frac{\Delta G_{(T)}}{T} - \frac{\Delta G^{\mathrm{o}}}{298} = \left(\frac{1}{T} - \frac{1}{298}\right)\Delta H^{\mathrm{o}} \tag{12.19}$$

The pressure dependence of G depends greatly on whether the system contains only liquids and solids or only gases. This can be shown easily if it is considered that ΔG for any pure material is simply the integral of Eq. (12.12) from the standard state (1 atm) at constant temperature (state 1 in Eq. (12.18)). If $dT = 0$, then Eq. (12.12) becomes

$$dG = V\,dP \tag{12.20}$$

Integrating gives

$$\int_{p_1}^{P_2} dG = \int_{p_1}^{P_2} V\,dP \tag{12.21}$$

Since state 1 is defined here as G^{o}, the following can be written:

$$G - G^{\mathrm{o}} = \int_{p_1}^{P_2} V\,dP \tag{12.22}$$

Again, using the identity for an ideal gas that $V = \dfrac{nRT}{P}$

$$G - G^{\mathrm{o}} = \int_{p_1}^{P_2} \frac{nRT}{P} dP \tag{12.23}$$

which evaluates to

$$G - G^{\mathrm{o}} = nRT \ln \frac{P_2}{P_1} \tag{12.24}$$

and finally

$$G = G^{\mathrm{o}} + nRT \ln \frac{P_2}{P_1} \tag{12.25}$$

This is the result for any gas that can be treated as an ideal gas.

For solids or liquids, however, the volume change with pressure under all but the most extreme circumstances is negligible, and volume can be removed from the integral in Eq. (12.22). This provides the following:

$$G = G_{\mathrm{o}} + V(P - 1) \tag{12.26}$$

However, on a molar basis, the volume of condensed phases is generally very small and so the expression $V(P-1)$ is quite small, and the following equation can be written:

$$G = G^{\mathrm{o}} \tag{12.27}$$

This says that, for condensed phases, G depends on temperature alone and not significantly on pressure.

12.3 The Free Energy Per Mole, μ, Is an Important Thermodynamic Quantity

In Eq. (12.23), the use of the ideal gas law led to the introduction of the mole quantity. This is an important consequence for questions that involve chemical processes, as most questions in biological science do. The relationship of the free energy on a per mole basis is so important that a special symbol, **μ**, is used to denote the relationship:

$$\mu_1 = \left(\frac{\partial}{\partial n_i}\right)_{T,P,n_{j\neq i}} \tag{12.28}$$

Here the term n_i represents the concentration of the component under consideration, and n_j represents all other components in the system. Substituting μ into Eq. (12.25) results in an expression describing the *molar free energy*:

$$\mu = \mu^{\mathrm{o}} + nRT\ln\frac{P_2}{P_1} \tag{12.29}$$

Since P_1 has already been defined as the standard state where $P_1 = 1$ atm, this equation becomes

$$\mu = \mu^{\mathrm{o}} + nRT\ln P \tag{12.30}$$

12.4 The Concept of Activity Relates an Ideal System to a Real System

So far, the point has been made and emphasized repeatedly that the equations being derived are general for any ideal substance. The point of fact is that few materials behave in an ideal manner, and especially, few systems that involve water and biological environments can be described in this ideal manner. It could seem that the efforts of the last chapters have been an exercise in futility. Fortunately, this is not the case. For the case of the ideal gas described by Eqs. (12.30) and (12.25), the free energy is related to its pressure. To continue to use the same general mathematics, which are simple and elegant, all that needs to be done is to invent a new state function that is related to the free energy of a real gas or any real substance. In the case of the gas, the new state function is the *fugacity*, ***f***, and substituting it into Eq. (12.30) gives the following:

$$\mu = \mu^{\mathrm{o}} + nRT\ln f \tag{12.31}$$

It is important that this new function, f, can be related to some real property of the gas. The ratio of fugacity to pressure, f/P, will give a fugacity coefficient γ.

At low pressure the fugacity will be equal to the pressure (i.e., $\lim_{P\to 0} f = P$) and when γ equals 1, the gas is ideal. When γ is less than 1 attractive intermolecular forces are dominant and when γ is greater than 1 repulsive intermolecular forces are dominant. Fugacity cannot be measured but must be calculated. Further discussion about fugacity is available in standard physical chemistry texts.

In the next sections, it will be shown that the free energy in an ideal solution is a function of concentration, and concentration can be treated just as pressure was for the ideal gas. Consequently, when the move is made from an ideal to a real solution, inventing a new state function will allow the same mathematics to be used for ideal gases and solutions and real gases and solutions. The new function in the case of the solution will be called the *activity* and given the symbol a.

The fortunate aspect of the ability to define a new state function as necessary is the powerful result that the relationships developed to this point are general and may be used in specific cases as long as the appropriate conditions and constraints are met and defined. It is now time to expand consideration toward aspects of systems that are relevant to biological study. The first expansion will be to consider the effects of a change in the components of a system as a change in state occurs.

12.5 The Application of Free Energy Considerations to Multiple-Component Systems

So far, the systems that have been considered have been implicitly defined as having a single component. In the study of biological systems, this is not a particularly useful simplification. Not only are most of the systems of concern comprised of a number of different materials, but frequently the amounts of these materials change as a system changes state because of chemical reactions that take place. Therefore, it now becomes necessary to consider explicitly the various components that make up a system and what their relationship will be to the free energy of the system.

For a pure substance, the free energy has been defined as

$$dG = V\,dP - S\,dT \tag{12.12}$$

Now consider a system that is comprised of a number of components present in different quantities, i.e., n_1, n_2, n_3,. . ., n_i. It is already apparent from the general case described in Eq. (12.28) that there is a dependence of G on the mole quantity. Therefore, for this expanded system of several components, what is of interest is G in terms of $G(T, P, n_1, n_2, n_3,. . ., n_i)$. The total differential is now

$$dG = \left(\frac{\partial G}{\partial T}\right)_{P,n_i} dT + \left(\frac{\partial G}{\partial P}\right)_{T,n_i} dP + \left(\frac{\partial G}{\partial n_1}\right)_{P,n_j\neq n_i} dn_1 + \left(\frac{\partial G}{\partial n_2}\right)_{P,n_j\neq n_i} dn_2 + \cdots \tag{12.32}$$

where $n_i$represents all components, and n_j represents all components except the one under consideration. This equation is simply an expanded and therefore more general case but simplifies if appropriately constrained; i.e., if $dn_1 = 0$ and $dn_2 = 0$, etc., then the following familiar equation results:

$$dG = \left(\frac{\partial G}{\partial T}\right)_{P,\,n_i} dT + \left(\frac{\partial G}{\partial P}\right)_{T,\,n_i} dP \tag{12.33}$$

This equation looks similar to Eq. (12.14) except that it applies to the case where the system contains different components provided the number of moles of each, n_i, is kept constant.

Equation (12.28) defined the molar free energy, μ. This is an expression that is similar to the new terms for the total differential equation in Eq. (12.32). Accordingly, as in the identities established for –*S* and *V*, the following will be defined:

$$\left(\frac{\partial G}{\partial n_i}\right)_{\mathrm{P,T},\,n_j} = \mu_i \tag{12.34}$$

Now, rewriting the total differential of $G(T, P, n_1, n_2, n_3,\ldots, n_i)$

$$dG = V\,dP - S\,dT + \mu_1 dn_1 + \mu_2 dn_2 + \cdots \tag{12.35}$$

is the same as writing

$$dG = V\,dP - S\,dT + \sum_i \mu_i dn_i \tag{12.36}$$

What has been accomplished here is significant. A relationship between the free energy of a system and the components that make it up has been described and given mathematical form. This quantity, μ, is called the *chemical potential,* and the property μ_i defines the chemical potential of substance *i*.

12.6 The Chemical Potential Is an Important Driving Force in Biochemical Systems

Chemical potential defines the property of matter that leads to the flow of a material from a region of high potential to a region of low potential just as gravity causes the movement of a mass from a region of high potential to one of lower potential. Looking at the partial derivative for μ

$$\left(\frac{\partial G}{\partial N_i}\right)_{P,T,n_j} = \mu_i \tag{12.37}$$

makes it clear that, as the molar quantity of an added material to a system is increased, the free energy is also increased. Following the addition, the material will seek to be distributed equally throughout the mixture so that no part of the volume that contains the mixture will be at a higher chemical potential than any other part. This is an important quality of the equilibrium condition when dealing with chemical potential. *At equilibrium, μ_i must have the same value throughout the system*. From our discussion we know that intensive variables are uniform throughout a system. Therefore, it is expected that μ_i would have the same value everywhere in the system at equilibrium.

Consider again

$$\mu = \mu^{\mathrm{o}} + RT \ln f \tag{12.31}$$

This was derived from the ideal statement

$$\mu = \mu^{\mathrm{o}} + RT \ln P \tag{12.30}$$

Because of the uniform properties of μ at equilibrium, it follows that the pressure and its related quantity fugacity are also uniform throughout the system at equilibrium. If a system of various (gaseous) components is considered, the chemical potential of each component will be given by

$$\mu_i = \mu_i^{\mathrm{o}} + RT \ln P_i \tag{12.38}$$

where $\mu_i^{\mathrm{o}}(T)$ is the chemical potential for a pure gas at temperature T and 1 atm of pressure. P_i is equivalent to writing the mole fraction of substance i at pressure P, i.e., $P_i = X_i P$. By substituting this identity for $P_i$and algebraically manipulating Eq. (12.32), a generalized relation can be found:

$$\mu_i = \mu_{i(T,P)}^{\mathrm{o}} + RT \ln X_i \tag{12.39}$$

This equation provides the result that a pure substance ($\mu_{i(T,P)}^{\mathrm{o}}$) in the same form (gas, liquid, or solid) as a mixture will flow spontaneously into the mixture. This provides the argument that different gases, liquids, and solids will diffuse one into the other.

12.6.1 Characteristics of μ

Several qualities of the chemical potential are worth highlighting. Consider a system in which both temperature and pressure are kept constant. Under these conditions, the free energy of the system will be determined solely by its composition. The following can be written:

$$G = \sum_i n_i u_i \tag{12.40}$$

This result characterizes the *additive* nature of the chemical potential. In the appropriate system at specified temperature and pressure, the free energy of the system can be computed if the molar quantities of the components are known.

Another important relationship of the chemical potential is derived from Eq. (12.40). If this equation is differentiated, the following result is obtained:

$$dG = \sum_i n_i du_i + \sum_i u_i dn_i \tag{12.41}$$

However, by Eq. (12.36), the following identity can be written

$$\sum_i n_i du_i + \sum_i u_i dn_i = V\,dP - S\,dT + \sum_i \mu_i dn_i \tag{12.42}$$

Solving the algebra yields

$$\sum_i n_i du_i = V\,dP - S\,dT \tag{12.43}$$

This is the *Gibbs–Duhem* equation, a special case of which may be written when pressure and temperature are constrained to be constant. Then, Eq. (12.43) reduces to

$$\sum_i n_i du_i = 0 \tag{12.44}$$

at constant *T*, *P*. This result makes it clear that there is a dependent relationship between the components making up a system. If a system is made up of several constituents and the chemical potential of one of them changes, then the others must also change coincidentally in some fashion, since the sum of the chemical potentials must remain zero.

Because the chemical potential seeks to be uniform throughout a system, an important result is found when equilibrium is prevented. Consider the effect of a barrier, such as a permeable membrane, on a chemical species whose μ on the left side is not equal to that on the right side. In this situation, there will be a force measured by ΔG that exists across the membrane. This force will appear as a driving force and will have the ability to perform work in the system. This is an extremely important consequence of the state function, μ, as will be seen in later chapters.

12.7 Entropy and Enthalpy Contribute to Calculations of the Free Energy of Mixing

The discussion so far has really been about the free energy change when one substance is mixed with another. Free energy, however, is the result of a relationship between the entropy and the enthalpy of a system. In considering the *free energy of* *mixing*, it will be of value to examine the constituent parts of the free energy of mixing, the entropy, and enthalpy of mixing. In this derivation, the model system will be a container that contains two pure substances that, in the initial state, are each separately found at constant T under pressure P. In the final state, the substances will be at the same T and the same P as found initially, but mixed together.

For the pure substances, $G_1 = n_1\mu_1^{\mathrm{o}}$ and $G_2 = n_2\mu_2^{\mathrm{o}}$. Therefore, the free energy of the initial state will be found by

$$G_{\text{initial}} = G_1 + G_2 = n_1\mu_1^{\mathrm{o}} + n_2\mu_2^{\mathrm{o}} = \sum_i n_i\mu_i^{\mathrm{o}} \tag{12.45}$$

In the final state, Eq. (12.40) will provide the value of G by adding together each part of $n_i\ \mu_i$, giving

$$G_{\text{final}} = \sum_i n_i\mu_i \tag{12.46}$$

ΔG_{mix} is equal simply to the initial value of G subtracted from the final value of G:

$$\Delta G_{\text{mix}} = \sum_i n_i\left(\mu_i - \mu_i^{\mathrm{o}}\right) \tag{12.47}$$

By Eq. (12.39), the term $\mu_i - \mu_i^{\mathrm{o}}$ can be evaluated and substituted as

$$\Delta G_{\text{mix}} = RT\,(n_1 \ln X_1 + n_2 \ln X_2) = \sum_i RTn_i \ln X_i \tag{12.48}$$

Sometimes this is more conveniently written by factoring out n_i by the relationship $n_i = x_i N$, where N is the total number of moles in the mixture, and X_i is the mole fraction of component i. This gives

$$\Delta G_{\text{mix}} = NRT\sum_i X_i \ln X_i \tag{12.49}$$

Solving this equation for a two-component mixture would show that the maximal free energy decrease occurs when equal amounts of two components are added to make the final mixture. In the case of four components, the maximal decrease will be for the case when the components each represents one-fourth of the final mixture, and so forth.

The *entropy of mixing* is found if ΔG is differentiated with respect to temperature:

$$\left(\frac{\partial\Delta G_{\text{mix}}}{\partial T}\right)_{P,n_i} = -\Delta S \tag{12.50}$$

Thus, differentiating Eq. (12.48) with respect to temperature

$$\left(\frac{\partial \Delta G_{\text{mix}}}{\partial T}\right)_{P,n_i} = NR\sum_i X_i \ln X_i \tag{12.51}$$

will give

$$\Delta S_{\text{mix}} = -NR\sum_i X_i \ln X_i \tag{12.52}$$

Since $X_i < 1$, this solution dictates that the entropy of mixing is always positive and the free energy of mixing is always negative. The argument that entropy always increases with mixing is reasonable in the statistical sense since there is far more choice in the mixed system, where many arrangements of different molecules look identical, as compared to the initial state, where there is only one possible arrangement of the different molecules. As in the solution of Eq. (12.49), the entropy of mixing is greater for a binary mixture if the components are each equal to half of the final mixture, and so on.

Finally, the *heat of mixing* can be found most easily simply by solving the fundamental equation:

$$\Delta G_{\text{mix}} = \Delta H_{\text{mix}} - T\Delta S_{\text{mix}} \tag{12.53}$$

Using the work already completed provides the result that

$$NRT\sum_i X_i \ln X_i = \Delta H_{\text{mix}} - T\left(-NR\sum_i X_i \ln X_i\right) \tag{12.54}$$

which reduces to

$$NRT\sum_i X_i \ln X_i = \Delta H_{\text{mix}} + NRT\sum_i X_i \ln X_i \tag{12.55}$$

and finally

$$\Delta H_{\text{mix}} = 0 \tag{12.56}$$

Therefore, for an ideal mixture, there is no heat associated with its formation. This holds true for any components of the ideal mixture whether the components are ions, atoms, or macromolecules. But what about a non-ideal mixture?

To answer this question, the following is considered. First, rewrite the fundamental equation for the free energy of mixing using the results to this point

$$-\Delta G = T\Delta S \tag{12.57}$$

The spontaneous process of mixing is wholly driven by entropy according to Eq. (12.57). If Eq. (12.52) is solved for its maximum (using a binary mixture, let x= 1/2), a value of approximately 5.76 JK^{-1}mol^{-1} is the result. For a biological system at 310 K, this gives a value for ΔG of –102.12 J. This is not a large number, and in non-ideal mixtures if the heat of mixing is more than a little positive, the two solutions will remain separate in immiscible layers.

12.8 Finding the Chemical Equilibrium of a System Is Possible by Making Free Energy Calculations

To this point, our discussion has focused on the question of mixing pure substances together. What happens when the substances in the initial state react to form different components in the final state? In other words, how does free energy relate to chemical reactions? Consider a simple chemical reaction

$$aA + bB\,\text{É}\,cC + dD \tag{12.58}$$

Because $\Delta G = G_{\text{final}} - G_{\text{initial}}$ or $\Delta G = G_{\text{products}} - G_{\text{reactants}}$, the following may be written:

$$\Delta G = (cG_C + dG_D) - (aG_A + bG_B) \tag{12.59}$$

This is the familiar statement of *Hess's law* that has been previously mentioned. It will be useful to be able to write ΔG in terms of a change in product–reactant behavior as was done earlier:

$$G = G^{\text{o}} + nRT\ln\frac{P_2}{P_1} \tag{12.25}$$

A look at Eq. (12.25), however, finds the equation in the form for an ideal gas. There is some advantage to this because in fact this equation is the general form for any ideal substance, whether gas, liquid, or solid. However, since the majority of the cases of interest for the remainder of this book will deal with solutions, it is a sensible side trip to derive a general form of Eq. (12.25) now so that it may be used henceforth.

12.8.1 Derivation of the Activity

In the discussion of the properties of the chemical potential for a given species, μ_i, we noted that at equilibrium, μ_i would be equal throughout a mixture. In the case of an ideal gas, this dictated that the partial pressure exerted by a component of the mixture would therefore be equal everywhere at equilibrium and that it would represent the free energy directly. In the case of a non-ideal or real gas, the pressure could be correctly substituted with the term fugacity, and the same relationship between the intensive property μ and the fugacity would exist as did for pressure and μ. The general form of the equation will hold true for any system, ideal or real, if a new term is introduced that will have the same properties with respect to μ (as does pressure or fugacity). This new quantity is called the *activity*, ***a***. If activity is substituted into Eq. (12.30),

$$\mu_i = \mu_i^{\mathrm{o}} + nRT \ln P_i \tag{12.30a}$$

which, it should be recalled, was derived from Eq. (12.25) directly the following new equation is the result:

$$\mu_i = \mu_i^{\mathrm{o}} + nRT \ln a_i \tag{12.60}$$

This is the molar free energy for a single component, *i*, in a system and, because of the identity of μ, may be rewritten as

$$\overline{G}_i = \overline{G}_i^{\mathrm{o}} + nRT \ln a_i \tag{12.61}$$

where $\overline{G}_i$ is the free energy per mole of component *i* in a mixture, $\overline{G}_i^{\mathrm{o}}$ is the standard free energy of a mole of component *i*, and a_i is the activity of component *i* in the mixture.

12.8.2 Activity of the Standard State

The activity of a substance in its standard state is easily found through the use of Eq. (12.61). This equation shows that the activity of a component is related to its standard state in the following manner:

$$\ln a_i = \frac{\overline{G}_i - \overline{G}_i^{\mathrm{o}}}{RT} \tag{12.62}$$

Since (ln 1) is zero, it follows that the activity of a material in its standard state is always equal to unity. This presents the question: What are the standard states for substances? For an ideal gas, the standard state is defined as a gas with a partial pressure equal to one. This is consistent with Eq. (12.25). For a real gas, the fugacity is given by a coefficient of activity, γ, multiplied by the partial pressure of the gas

$$f = \gamma_i P_i \tag{12.63}$$

At pressures approaching zero, γ becomes equal to unity. It is for this reason that in low-pressure systems a real gas may be treated as an ideal gas, so then the pressure directly indicates the fugacity. For a pure solid or liquid, the activity is unity at 1 atm of pressure. Because the free energy of a solid has been shown to be essentially independent of pressure (Eqs. (12.26) and (12.27)), it is generally correct to write that at any pressure the activity of a pure solid or liquid is unity.

The standard state of a solution is a little more difficult to define than that of a pure substance. Since in Part III a great deal of attention will be focused on solutions, let it be sufficient at this point to indicate that the activity of a solution depends on an activity coefficient and the concentration of the component in the solution. One of the problems in defining a standard state arises from the inclusion of the concentration. There are quite a few methods of measuring concentration in solutions, and hence confusion can be easily generated. As an example of the general form for the activity of a component, *i*, of a solution, the specific formula for the mole fraction and the solvent standard state will be written as

$$a_i = \gamma X_i \tag{12.64}$$

Here γ is the activity coefficient for this standard state, and X_i is the mole fraction of component *i*. Mole fractions are a convenient measure of concentration and represent the number of moles of component *i* in a mixture divided by the total number of moles found in the mixture. As the value of X_i approaches unity, the number of moles of component *i* becomes closer and closer to the total number of moles that comprise the entire system. The activity coefficient is such that when the mole fraction X_i is equal to unity, the activity of component *i* in the solution is also unity. This is a particularly valuable quality, since when a solution is a single pure component – in other words, the initial state of the solvent itself – it will have an activity of one. This result is in agreement with the result for the standard state of a pure liquid already given.

12.8.3 The Equilibrium Expression

Now, the problem of defining the free energy for Eq. (12.59) can be addressed. Using Eq. (12.61), which defines *G* in terms of activity, the following can be written:

$$\begin{aligned} \Delta G =& \left(c\overline{G}^{\mathrm{o}}_{C} + d\overline{G}^{\mathrm{o}}_{C}\right) - \left(a\overline{G}^{\mathrm{o}}_{A} + b\overline{G}^{\mathrm{o}}_{B}\right) \\ =& \left(cRT\ln a_c + dRT\ln a_D\right) - \left(aRT\ln a_A + bRT\ln a_B\right) \end{aligned} \tag{12.65}$$

The terms can be collected and Eq. (12.65) can be written in the more convenient form as

$$\Delta G = \Delta G^{\mathrm{o}} + RT\ln\frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.66}$$

This equation relates the free energy change found when reactants and products with any arbitrary value of activities are mixed together or are formed by a chemical reaction. The change in G is taken relative to the standard state and so the symbol G^{o} is used. The ratio of activities in Eq. (12.66) is called Q:

$$Q = \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.67}$$

Q is the ratio of the activities of the components each raised to the power of its coefficient in the chemical reaction. Changing mole numbers arithmetically will change Q geometrically; that is, doubling the mole number will square Q. When the product activities are large relative to the reactant activities, Q will be large and its contribution to the free energy of the reaction will be positive, thus making the reaction less likely to occur spontaneously. Conversely, a small Q will result in a negative contribution to the free energy, and thus will make the reaction more likely to occur. Q is not necessarily representative of a system at equilibrium, but such a case can be considered.

At equilibrium, the ratios of products and reactants are given the special designation K, the equilibrium constant:

$$K = \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.68}$$

The equilibrium constant K is a special case of Q, and in this case, $\Delta G = 0$. Using the equilibrium constant in place of Q

$$0 = \Delta G^{\mathrm{o}} + RT\ln K \tag{12.69}$$

which is equal to

$$\Delta G^{\mathrm{o}} + RT\ln K \tag{12.70}$$

Several other algebraic manipulations provide convenient formulas:

$$K = e^{-\Delta G^{\mathrm{o}}/RT} \tag{12.71}$$

and

$$K = 10^{-\Delta G^{\mathrm{o}}/2.303RT} \tag{12.72}$$

All of these equations are valuable because they allow calculation of the standard free energy change for a reaction simply by measuring the amounts of products and reactants in a system. This is precisely the kind of information that is valuable in biochemical studies.

Implicit in this discussion is that the free energy of a system depends on the concentration of the components that make up the system. Changes in the concentration of various products and/or reactants can lead to the spontaneous progression of a reaction forward or backward. When the activity of various components is changed, the system responds by attempting to move toward a new equilibrium position. Such a response is the same as that predicted by the Le Chatelier principle.

At this point, two clear methods exist to predict the movement of a system in one direction or another depending either on the "natural" tendencies or on work being done on or by the system. The experimenter may either use the equilibrium constant to predict the free energy change or use the equation

$$\Delta G = \Delta H - T\Delta S \tag{12.73}$$

and solve for ΔG. Enthalpy may be obtained either from a table (Table 3.7) and Hess's law or from careful thermometry and calorimetry using $\Delta H = q$. ΔS may also be found using Hess's law, or by careful thermal measurements of heat capacities on pure substances. A third and very valuable method of finding ΔG is also available and is an electrochemical method. In many cases, this method is preferred, and its derivation and use will be considered following a discussion of the temperature dependence of the equilibrium constant.

A very useful relationship allows the determination of entropy and enthalpy changes for a reaction by measuring the temperature dependence of the equilibrium constant. Consider the equilibrium constant for each of two temperatures, T_1 and T_2

$$\ln K_{(T_1)} = \frac{-\Delta G^{\mathrm{o}}_{(T_1)}}{RT_1} \tag{12.74}$$

and

$$\ln K_{(T_2)} = \frac{-\Delta G^{\mathrm{o}}_{(T_2)}}{RT_2} \tag{12.75}$$

Their relationship can be found by subtracting one from the other

$$\ln K_{(T_2)} - \ln K_{(T_1)} = \frac{-\Delta G^{\mathrm{o}}_{(T_2)}}{RT_2} - \left[\frac{-\Delta G^{\mathrm{o}}_{(T_1)}}{RT_1}\right] \tag{12.76}$$

This equation can be rearranged so that for ln K

$$\ln K_{(T_2)} = \ln K_{(T_1)} - \left[\frac{\Delta G^{\mathrm{o}}_{(T_2)}}{RT_2} - \frac{\Delta G^{\mathrm{o}}_{(T_1)}}{RT_1}\right] \tag{12.77}$$

Using $\Delta G = \Delta H - T\Delta S$ and rewriting Eq. (12.77) in terms of entropy and enthalpy

$$\ln K_{(T_2)} = \ln K_{(T_1)} - \left[\frac{\Delta H^{\mathrm{o}}_{(T_2)}}{RT_2} - \frac{\Delta H^{\mathrm{o}}_{(T_1)}}{RT_1} - \frac{\Delta S^{\mathrm{o}}_{(T_2)}}{R} + \frac{\Delta S^{\mathrm{o}}_{(T_1)}}{R}\right] \tag{12.78}$$

Assuming that ΔH and ΔS do not change significantly over the temperature range under consideration, Eq. (12.77) can be simplified:

$$\ln K_{(T_2)} = \ln K_{(T_1)} - \left[\frac{\Delta H^{\mathrm{o}}_{(T_1)}}{R} \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \right] \tag{12.79}$$

$$\ln K_{(T_2)} = \ln K_{(T_1)} + \left[\frac{\Delta H^{\mathrm{o}}_{(T_1)}}{RT_1T_2} (T_2 - T_1) \right] \tag{12.80}$$

Equation (12.80) indicates the shift in equilibrium with a variation in temperature. If a reaction is exothermic ($\Delta H < 0$), then an increase in temperature will favor the reactants; but if the reaction is endothermic ($\Delta H > 0$), an increase in temperature will favor product formation.

An exact result can be found for cases where the approximations leading to Eq. (12.78) are not valid, by using the Gibbs–Helmholtz equation (Eq. (12.17)):

$$\left(\frac{\partial \left({}^{G}/_{T} \right)}{\partial T} \right)_{\mathrm{P}} = -\frac{H}{T^2} \tag{12.17}$$

which gives the van’t Hoff equation

$$\frac{d \ln K}{dT} = -\frac{\Delta H^{\mathrm{o}}_{(T)}}{RT^2} \tag{12.81}$$

Graphing $\ln K$ versus $1/T$ gives a slope of $-\frac{\Delta H^{\mathrm{o}}}{R}$. Therefore, by measuring equilibrium concentrations over a range of temperatures and using the van′t Hoff equation, the enthalpy of a reaction may be found without resorting to calorimetry.

12.9 The Thermodynamics of Galvanic Cells Is Directly Related to the Gibbs Free Energy

A system consisting of an electrolyte and two electrodes, with the two electrodes connected together through an external electric circuit, is called a *galvanic cell*. When a chemical process occurs in the cell that causes electricity to flow, the flow of electricity can be sensed by a voltmeter placed in the external circuit of the cell. The potential that is established depends directly on the free energy change of the process. If appropriate electrodes are chosen that are sensitive to a given process under study, a galvanic cell can be used to determine ΔG at equilibrium. A galvanic cell can be used in the study of, for example, oxidation–reduction reactions, concentration differences of a species separated by a membrane, formation of complex ions, acid–base dissociations, and solubilities of salts. For a galvanic cell to measure ΔG, the cell must be acting reversibly; that is, the application of an infinitesimal potential across the cell will move the reaction equally in one direction or another. Stated another way, if the resistance of the external measuring circuit is ∞, no current will flow in the external circuit. The voltage measured between the two electrodes then represents the free energy change exactly. All of the energy resulting from such a system is theoretically available to do work. We know that electrical work is given by the product of charge and potential:

$$w_{\text{electrical}} = -QE \tag{12.82}$$

This can be rewritten as

$$w = -nF\varepsilon \tag{12.83}$$

where ε is equal to E in volts, and Q is written as the product of the electrical equivalents per mole, n, and the faraday, F. The faraday, F, is the charge associated with 1 mole of electrons or 96,487 C mol^{-1}. The sign is negative because the reversible work is done by the system on the surroundings and reduces the free energy of the system. Therefore, for a system at constant temperature and pressure

$$\Delta G = -nF\varepsilon \tag{12.84}$$

If the reaction in the cell is given by

$$aA + bB \rightleftharpoons cC + dD \tag{12.58}$$

then the free energy change for the reaction is given by Eq. (12.65)

$$\Delta G = \Delta G^{\text{o}} + RT \ln \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.66}$$

If Eqs. (12.66) and (12.84) are equated, then the following equation results

$$-nF\varepsilon = \Delta G^{\text{o}} + RT \ln \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.85}$$

Dividing by $-nF$ gives the potential for any reaction occurring that is not in the standard state:

$$\varepsilon = \frac{-\Delta G^{\text{o}}}{nF} - \frac{RT}{nF} \ln \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.86}$$

A standard emf potential, ε^{o}, for a cell in which the components are in a defined standard state can be written as

$$\varepsilon^{\text{o}} = \frac{-\Delta G^{\text{o}}}{nF} \tag{12.87}$$

The standard state is defined when the activity of each of the components is equal to unity. The biochemist′s standard state for ε^{o} is at constant pressure of 1 atm, 298 K, and pH = 7. Equation (12.87) is the same as the first term in Eq. (12.86), and substitution gives

$$\varepsilon = \varepsilon^{o} - \frac{RT}{nF} \ln \frac{(a_C)^c\,(a_d)^d}{(a_A)^a\,(a_B)^b} \tag{12.88}$$

This is the *Nernst equation* and relates the activity of the components in the system comprising the galvanic cell to the standard state emf, ε^{o}. The Nernst equation will be the starting point for the discussion of membrane potentials in Chapter 24. ε^{o} is significant since at equilibrium it can be written in terms of K, i.e.,

$$\varepsilon^{o} = -\frac{RT}{nF} \ln K \tag{12.89}$$

which relates the standard cell emf to the equilibrium constant of the cell reaction.

Knowing the temperature dependence of the emf allows calculation of ΔH^{o} and ΔS^{o}. If Eq. (12.83) is written for the standard state

$$\Delta G^{o} = -nF\varepsilon^{o} \tag{12.90}$$

and this equation is differentiated with respect to temperature, the following is the result:

$$\left(\frac{\partial \Delta G^{o}}{\partial T}\right)_{\mathrm{P}} = -nF\left(\frac{\partial \varepsilon^{o}}{\partial T}\right)_{\mathrm{P}} \tag{12.91}$$

Because $(\partial \Delta G^{o}/\partial T)_{\mathrm{P}}$ is equal to $-\Delta S^{o}$, the following can be written as

$$\Delta S^{o} = nF\left(\frac{\partial \varepsilon^{o}}{\partial T}\right)_{\mathrm{P}} \tag{12.92}$$

Using the relation for enthalpy, $\Delta H^{o} = \Delta G^{o} + T\Delta S^{o}$, the enthalpy, ΔH^{o}, can be found by

$$\Delta H^{o} = -nF\varepsilon^{o} + nFT\left(\frac{\partial \varepsilon^{o}}{\partial T}\right)_{\mathrm{P}} \tag{12.93}$$

The activity of a component can also be measured directly with an electrode system in terms of the mean ionic activity, $a_{\pm}$, and mean ionic molality, $m_{\pm}$. The activities of components are explicitly described by the Nernst equation, and if ε is measured at a particular concentration of an electrolyte, m, then the mean ionic activity of the electrolyte can be determined if ε^{o} is also known. The concept of the mean ionic activity will be discussed in Chapter 16.

12.10 Free Energy Changes Relate the Equilibrium Position of Biochemical Reactions

One of the most important results of the treatment of chemical equilibria by the Gibbs free energy is the ability to understand how coupling an exothermic reaction such as the hydrolysis of ATP can drive an otherwise non-spontaneous process. This coupling is used extensively in biochemical processes starting with the use of the field energy of the Sun to drive electron transfer reactions in plants. The production of high energy intermediates such as NADH, FADH, and ATP allow the free energy of one reaction to be directed into another reaction to perform necessary biological work. This work takes all manner of forms from information processing at the nerve cell (which requires the membrane potentials maintained by the ATP driven Na^+–K^+ ion pumps) and the energy-dependent repair of DNA mutations caused by exposure to UV light. Table 12.1 lists a few of the energy-coupled reactions in biological systems. Tables of the standard free energy changes that accompany biological reactions are widely available. Electrode potentials for the redox reactions that occur in biological systems can be used to calculate free energy changes for the reactions thus doubling the value of tables such as Table 3.7.

Table 12.1 Listing of selected coupled biochemical reactions

Reaction	$\Delta G^{o\prime}$ kJ/mol
Glucose + ATP→ glucose-6-phosphate + ADP	–14.2
Fructose-6-phosphate + ATP → fructose-1,6-diphosphate + ADP	–14.2
1,3-Diphosphoglycerate + ADP + H^+ → 3-phosphoglycerate + ATP	–28.5
2-Phosphenolpyruvate + ADP + H^+ → pyruvate + ATP	–23.8
Pyruvate + NADH + H^+ → lactate + NAD^+	–25.1

Free energies of hydrolysis for high energy phosphorylated compounds

Compound	$\Delta G^{o\prime}$ kJ/mol
Phosphoenolpyruvate	−62
Acetyl phosphate	−43
Creatine phosphate	−43
Pyrophosphate	−33
ATP	−30.5
Glycerol 6-phosphate	−9.2

Further Reading

Biological Free Energy

Urry D. (1997) Physical chemistry of biological free energy transduction as demonstrated by elastic protein-based polymers, *J. Phys. Chem.*, **101**:11007–11028.

Biosensors

An important application of chemical equilibrium and the electrochemical cell is the biosensor. The following provide an introduction to the field.

Light T.S. (1997) Industrial use and applications of ion selective electrodes, *J. Chem. Educ.*, **74:**171–177.

Frant M.S. (1997) Where did ion selective electrodes come from? The story of their development and commercialization, *J. Chem. Educ.*, **74:**159–166.

Ruzicka J. (1997) The seventies – golden age for ion selective electrodes, *J. Chem. Educ.*, **74:** 167–170.

Schultz J.S. (1991) Biosensors, *Sci. Am.*, **265, 2**:64–69.

Young C.C. (1997) Evolution of blood chemistry analysers based on ion selective electrodes, *J. Chem. Educ.*, **74:**177–182.

Coupled Reactions

Spencer J.N. (1992) Competitive and coupled reactions, *J. Chem. Educ.*, **69**:281–284.

Wink D. (1992) The conversion of chemical energy, technological examples, *J. Chem. Educ.*, **69**:108–110.

Wink D. (1992) The conversion of chemical energy, biochemical examples, *J. Chem. Educ.*, **69**:264–267.

Electrochemical Equilibria

Runo J.R. and Peters D.G. (1993) Climbing a potential ladder to understanding concepts in electrochemistry, *J. Chem. Ed.*, **70:**708–711. (An article bridging the gap from the concepts of equilibrium electrochemical cell reactions to dynamic electrochemistry.)

Problem Sets

1. The proton gradient generated by the electron transport chain in the mitochondria is approximately 1.5 pH unit with the outside lower than the inside pH of 7.0. There is a transmembrane potential of –140 mV, inside to outside. The production of ATP from ADP is driven by coupling the collapse of the electrochemical gradient to the formation of the pyrophosphate bond: ADP + $P_i \rightarrow$ ATP, $\Delta G^{o\prime} =$ 31 kJ mol^{-1}. The conditions inside the mitochondrion are pH = 7.0, $T = 310$ K, [ATP] = 1.5 mM, [P_i]= 3 mM and [ADP] = 1.5 mM.

   (a) Under these conditions what is the ΔG for the production of ATP?
   (b) What is the electrical energy that must be coupled to drive this reaction forward?

2. Show how measuring the voltage associated with a galvanic cell will give the equilibrium constant of the reaction.

3. The reduction potential of H_2S is –230 mV while that of $NADH_2$ is –320 mV (versus SHE). What color photon would be necessary to allow the coupled reaction of H_2S oxidation and $NADH_2$ reduction to proceed. Assume 100% quantum efficiency in the photochemical reaction.
4. On a cold winter morning, the difficulty in getting a car engine to start is often attributed to the cold temperature decreasing the "cranking power" of the battery. What is the actual reduction in cell voltage when the temperature is 260 K versus 298 K?
5. Add a column to Table 3.7 that lists the $\Delta G^{o\prime}$ for each of the reactions listed. Compare it with a reference table as is found in the *CRC Handbook of Chemistry and Physics*.

Chapter 13
The Thermodynamics of Phase Equilibria

Contents

13.1 The Concept of Phase Equilibrium Is Important in Biochemical Systems

So far we have considered only homogeneous solutions. Such systems, while composed of a number of constituents, contain them in a single homogeneous *phase*. A phase is considered to be a state in which there is both chemical and physical uniformity. In a biological system, such a presumption is not realistic. Biological systems have many heterogeneous characteristics. Many of the processes in cells and living organisms involve the transfer of chemical species from one phase to another. For

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_13, © Springer Science+Business Media, LLC 2010

example, the movement of ions across a membrane such as the cell membrane or an intracellular organelle is often treated as transport between two phases, one inside and the other outside. The nature of the equilibria that can exist between phases will be the focus of this chapter. When different phases come in contact with each other, an interface between them occurs. This interface is a surface and the properties of a surface are different from those of either of the phases responsible for creating it.

The second law of thermodynamics says that the chemical potential of each of the components is equal everywhere in any system in equilibrium. If two or more phases are in equilibrium, then the chemical potentials of each species in each phase of the system are equal. This can be written as

$$\mu_{\mathrm{i}}(\text{phase1}) = \mu_{\mathrm{i}}(\text{phase2}) = \ldots \tag{13.1}$$

An equivalent relationship can be written for each component or species of the system. Since the chemical potential depends on the activity of each species in each phase, Eq. (13.1) can be written as

$$\mu_{\mathrm{i}} = \mu_{\mathrm{i}}^{\mathrm{o}} + RT \ln a_{\mathrm{i(phase\ 1)}} = \mu_{\mathrm{i}}^{\mathrm{o}} + RT \ln a_{\mathrm{i(phase\ 2)}} \tag{13.2}$$

If the standard states of each component are equivalent, this equation expresses the important result that the activity of the species will be the same in each phase. Since the activity is defined as

$$a_i = \gamma_i\,[i] \tag{13.3}$$

where $[i]$ is the concentration of species i, this result means that, if the activity coefficients, γ, for the component are the same in each phase, the concentrations in each phase must also be equal. This is an important principle in biochemical research for it forms the basis of equilibrium dialysis studies. Although this principle of phase equilibrium has great practical significance in biological studies, it must be used with the greatest caution in approaching cellular problems. The activity coefficient, γ, is a formal term that represents the deviation from ideality of a particular component and is determined largely by the environment of the molecule and its interaction with that environment.

In a cellular system, the activity coefficient may reasonably be considered to always deviate from unity and hence the activity of virtually all species will vary considerably from the ideal. Many phases in cellular systems are separated by membranes and the resulting compartmentalization leads to environments that can differ radically. It is therefore generally difficult to presume that the activity coefficients in each phase are identical. In Section 13.3, some of the forces that are reflected in the activity coefficient will be discussed, primarily from the viewpoint of the solute. It is crucial to recognize that a common assumption about cellular chemistry is that, due to the small size and high molar concentration of liquid water, its activity can generally be taken as unity and its concentration as unchanging. The actual activities of water in the multiple compartmentalized microenvironments of the cell are presently unknown. However, it is probably reasonable to assume that the activity coefficients of all components, including the aqueous solvent, deviate from unity.

13.2 Thermodynamics of Transfer Between Phases

The function and survival of a cell and of the organisms comprised by cells are dependent to a great extent on the uptake and elimination of various molecules into and out of the intracellular environment. Inside most cells (with the exception of the red blood cell), the environment is dominated by membrane-defined sub-compartments, thus generating multiple intraorganelle phases that are related to the cytoplasmic phase. It is therefore often convenient to consider the cell (or a portion thereof) as a two-phase system with a membrane separating the phases. At constant temperature and pressure, the free energy change associated with the transfer of components between the phases depends on the chemical potential of each component:

$$\Delta G = \mu_{(\text{phase 2})} - \mu_{(\text{phase 1})} \tag{13.4}$$

or

$$\Delta G = RT \ln \frac{a_{(\text{phase2})}}{a_{(\text{phase1})}} \tag{13.5}$$

In the case where activity coefficients are equal in the different phases, then Eq. (13.5) may be written in terms of concentration alone. However, as discussed above, it is generally wiser to use Eq. (13.5) and explicitly express the activity coefficient. Equation (13.5) is adequate for determining the free energy of transfer between phases as long as the molecule is uncharged. If a molecule or ion has an electric charge and is in the presence of an electric field (as is almost always the case in cells), then a further term must be added to the free energy expression:

$$\Delta G = RT \ln \frac{a_{(\text{phase2})}}{a_{(\text{phase1})}} + nF\varepsilon \tag{13.6}$$

where n is the valence number of the ion, F is the faraday, and e is the potential field in volts. The potential field is expressed with reference to phase 1 and therefore the sign of e is positive if phase 2 is at a higher potential than phase 1.

13.3 The Phase Rule Relates the Number of Variables of State to the Number of Components and Phases at Equilibrium

The systems described so far in this chapter have been more complex than those considered earlier, notably because they have been said to consist of several phases. The reflective reader will be thinking at this point, “Just what is a phase and how is it defined anyhow? And how does one know just how many phases are present in a system?” In the analysis necessary to answer these questions, another very important question will also be answered: What are the constraints that apply to the thermodynamics of a system comprised of multiple phases? In other words, how many degrees of freedom exist that can change the variables of state in a system that is composed of a number of components and a number of phases at equilibrium? The solution to this problem was found in the nineteenth century by J. Willard Gibbs, the father of chemical thermodynamics, and stands as one of the most elegant results in the whole of modern science. The *phase rule* states that the maximum number of phases that can exist at equilibrium in the absence of external fields is equal to the number of components plus two. For a one-component system such as water, if external forces such as gravitational and electric fields are ignored, three phases may exist at equilibrium: ice, water, and steam. However, nature dictates that this three-phase equilibrium may only be achieved at the expense of the freedom to change the variables of state. If a single component such as water exists in a three-phase equilibrium, the state variables are defined; that is, the temperature and pressure of the system are constrained and may not be varied, as indicated by the result of zero obtained from the following equation:

$$F = 3 - P \tag{13.7}$$

where F is the number of degrees of freedom, ie, temperature and pressure, that can be varied, and P represents the number of phases that exist in equilibrium. As will be described later in this chapter, this point is called the *triple point*. Water can exist in numerous forms of crystalline ice at various temperatures and pressures, but the phase rule says that only a total of three of these phases, including water and steam, can exist together at equilibrium.

The more general case is considered for a system of more than one component. Again ignoring external fields, the phase rule is written in its classical form:

$$F = C - P + 2 \tag{13.8}$$

where C represents the number of components of the system. Typically, the variables considered are the temperature, the pressure, and the chemical potential of the components. For example, if a two-component system of NaCl and water is considered, Eq. (13.8) can be used to determine the maximum number of possible phases. The maximum number of phases occurs when the number of degrees of freedom to manipulate the system variables is minimum, i.e., $F = 0$:

$$\begin{aligned} 0 &= 2 - P + 2 \\ P &= 4 \end{aligned} \tag{13.9}$$

These phases might include solid NaCl, liquid electrolyte solution, ice, and water vapor. If, on the other hand, only three phases existed in equilibrium, i.e., solid NaCl, liquid electrolyte, and water vapor, a single degree of freedom would be available. Consequently, this three-phase system could exist in equilibrium at a range of temperatures or pressures. For each external force (gravitational, magnetic, electrical) imposed on the system, the constant 2 in Eq. (13.8) is increased by one.

This discussion has answered the reflective reader's questions except for how a phase is defined or identified. This in fact is usually simple but being warned that with a little imagination it is possible to be tricked and get into the thermodynamic quicksand. Gibbs himself wrote "a phase…is a state of matter that is uniform throughout not only in chemical composition but also in physical state." A homogeneous solution of NaCl is a single phase since it is a uniform mixture (on a reasonable time-averaged scale) of Na^+ and Cl^- ions and water molecules. Any reasonable or thermodynamic sample of the solution will accurately represent the whole solution. What are the gray areas for the application of the phase rule? Unfortunately for the biological worker, a microscopic examination of a *dispersion*, a material that is uniform on a macroscopic scale but is made up of distinct components embedded in a matrix, is a very gray area for applying the phase rule. Another gray area is consideration of what constitutes a phase in a gravitational or electric field. If a biological membrane is considered, the description of a dispersion fits it well; in addition the presence of strong electric fields is a well-proven fact. Care must be taken therefore when applying the phase rule in cellular biological systems.

13.4 The Equilibrium Between Different Phases Is Given by the Clapeyron Equation

A pure substance such as water will have equilibrium relationships with its various phases, solid, liquid, and vapor. The equilibria of these phase phenomena depend on the state variables. Traditionally, the variables that are related to the phase changes of a pure substance are temperature and pressure. These relationships are important in biological systems. The approach to phase transitions of a pure substance is useful in understanding the changes in conformation of macromolecules such as proteins or DNA. The equilibrium between the phases of a pure substance is also important because it forms the basis for defining the behavior of ideal solutions, the activity of solvents, osmotic pressure, and colligative properties of solutions. It must be noted that while phase transitions are most formally described in terms of temperature and pressure variations, the phase transition will be affected if other forces are present, for example, electric or magnetic fields.

The equilibrium between phases of a pure substance is given by the *Clapeyron equation*:

$$\frac{dP}{dT} = \frac{\Delta S}{\Delta V} \tag{13.10}$$

This equation is derived from the condition that

$$\mu(T,P)_a = \mu(T,P)_b \tag{13.11}$$

where a and b represent two phases in equilibrium. Writing this in terms of the fundamental equation ($\mu = -SdT + VdP$) gives

$$(-SdT + VdP)_a = (-Sdt + VdP)_b \tag{13.12}$$

which is

$$(S_b - S_a)dT = (V_b - V_a)dP \tag{13.13}$$

The expressions in parentheses are ΔS and ΔV, respectively, and therefore Eq. (13.13) can be written as

$$\Delta S dT = \Delta V dP \tag{13.14}$$

Rearrangement gives Eq. (13.10), the Clapeyron equation. The Clapeyron equation describes the relationship between the equilibrium temperature and pressure quantitatively. Phase diagrams of the equilibrium pressure against temperature can be constructed using this equation. A phase diagram for water, for example, could be constructed by applying the Clapeyron equation to the phase transitions between solid and liquid, liquid and gas, and gas and solid states.

Consider first, the biochemically relevant case of water at moderate pressure. The terms ΔS and ΔV are given by ΔS_{fusion} and ΔV_{fusion}. At the equilibrium temperature, ΔS_{fusion} is reversible and is equal to

$$\Delta S_{\text{fusion}} = \left(\frac{\Delta H_{\text{fusion}}}{T}\right) \tag{13.15}$$

Heat is always absorbed as the state changes from solid to liquid or gas ($\Delta H_{\text{fusion}} > 0$). ΔS_{fusion} is therefore positive for all substances. For all substances undergoing phase transformations from solid to gas (sublimation) and liquid to gas (vaporization) as well as solid to liquid, ΔH is greater than 0 and therefore ΔS for any of these phase transitions is always positive. The term ΔV_{fusion} will be positive or negative depending on whether the density of the solid or the liquid phase is greater. For most compounds, the solid phase is slightly more dense than the liquid phase, so ΔV_{fusion} is positive and hence $\Delta S_{\text{fusion}}/\Delta V_{\text{fusion}}$, the slope of the line for the solid–liquid phase transition, is positive. In the case of water, however, the density of ice is less than that of liquid water and the slope of the line describing the phase transition between ice and water is negative (see Fig. 13.1). The variation in density between liquid and solid is usually small, and consequently, the slope of the line described for this phase transition is generally quite steep.

Similar lines may be drawn on the graph of Fig. 13.1 for the other phase transitions. In the case of liquid–gas phase transitions, the value of $\Delta H_{\text{vaporization}}$ is always greater than 0 since the density of a gas is less than that of a liquid in every case. The change in volume is generally quite significant and so the slope of the line will be small compared to that of the solid–liquid transition. Finally, a line may be drawn on the phase diagram representing the equilibrium relationship between solid and gas phases. This line also will have a positive slope. The phase diagram for water shows there are a series of equilibrium points described by the line between two phases, where ice and water vapor, ice and liquid water, and liquid water and water vapor are in equilibrium. Since the points that comprise the lines represent points

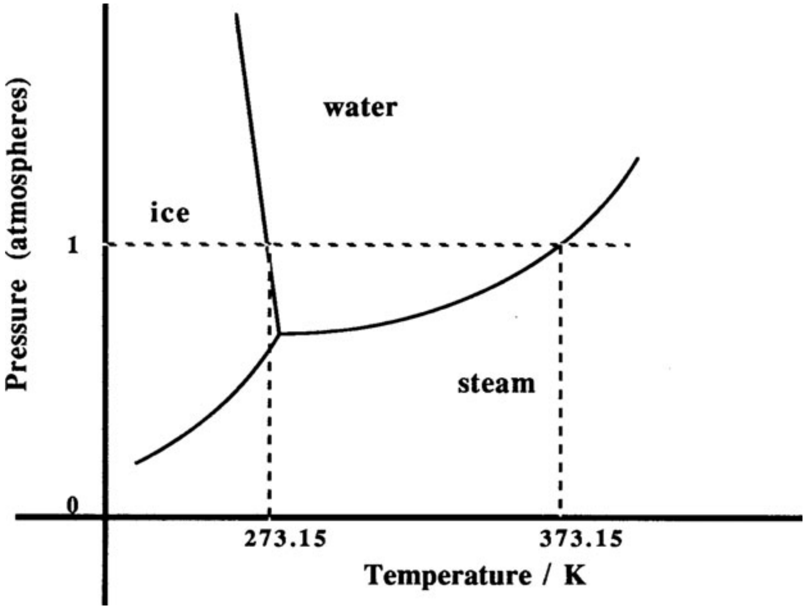


Fig. 13.1 Phase diagram of water. A phase diagram represents the variety of states that a system may have, constrained by the variables on the axes. In this figure, the state of a system of water is defined for each *x*-, *y*-coordinate representing a particular value of temperature and pressure. The *lines* on the graph define the special case where an equilibrium exists between the various phases of the water. For example, there are numerous states defined by a particular temperature or pressure where water exists as either a gas or a liquid. These states have *x*-, *y*- coordinates either above or below the line on the diagram. The systems defined at the points which fall on the line separating the gas and liquid states are the equilibrium states where both gas and liquid may coexist

of equilibrium between two phases, a single phase is represented by a point on the graph that does not fall on the line. At the intersection of the three lines, a unique state is described where all three phases of a substance may coexist. This is the triple point. Only one triple point can exist for a specific set of three phases. The phase diagram of water has been used to explain the thermodynamics involved in the sport of ice skating. It can be seen that the added pressure on the ice from the skate alters the local phase equilibria and the ice melts, providing a lubricating film of water. However, if the ambient temperature is too low, even the added pressure is not sufficient to generate the lubricating phase of water and the skater will stick to the ice surface and fall. This view has been widely taught, but it has been pointed out recently that the calculations do not hold up. Examination of the surface properties of ice has shown instead that the surface layer is water like and ice is intrinsically slippery because of a self-lubricating film of liquid-like water at the surface of ice. Thus the phase diagram at the surface of ice must include a liquid-like layer that under pressure and in cold temperatures changes the surface properties, not the bulk properties of the material.

The Clapeyron equation can also be used to predict the effect of a change in one system variable such as pressure on another, eg, the phase transition temperature. This has practical value in predicting changes in the boiling point ($T_{\text{vaporization}}$) at various pressures and is applied in pressure cookers and autoclaves. By integrating the Clapeyron equation, expressions can be derived for solid–liquid and condensed phase–vapor phase equilibria. These are as follows for cases of solid-to-liquid transition:

$$\Delta P = \frac{\Delta H_{\text{fusion}}}{\Delta V_{\text{fusion}}} \frac{\Delta T}{T_{\text{melt}}} \tag{13.16}$$

Another useful expression is the *Clausius–Clapeyron equation*. It is useful for cases of condensed phase–gas equilibria:

$$\ln \frac{P_2}{P_1} = \frac{-\Delta H_{\text{vaporization}}}{R} \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \tag{13.17}$$

It is important to recognize that the equation assumes that $\Delta H_{\text{vaporization}}$ is independent of temperature. Another derivation of this equation is given in Appendix L in which the relationship of the activity of the phases is discussed more explicitly. The Clausius–Clapeyron equation quantitates the increase in the activity of the substance in the condensed phase with respect to the pressure in the system. For example, the activity of oxygen in the body tissues is increased under hyperbaric conditions that change the equilibria of oxygen binding in numerous enzyme systems and in oxygen-binding proteins such as hemoglobin. The concept behind hyperbaric oxygen treatment for diseases, such as severe anaerobic infections and sickle cell anemia, is based on the presumed beneficial increase in oxygen activity.

As mentioned earlier, a very important aspect of phase equilibria in biological science is the use of the colligative properties (vapor pressure, osmotic pressure, boiling point, and freezing point) of a solution. The relationship of the colligative properties of a solution to the solute concentration is used to define an ideal solution and also to explicitly measure the activity of components in a solution. Such measurements can be used in biological applications for purposes that include the determination of the molecular weight of a macromolecule and of the osmolarity of a patient's serum.

13.4.1 Colligative Properties Vary with Solute Concentration

The colligative properties of a solution vary with the amount of solute. The boiling point is increased and the freezing point is depressed as the amount of solute added into solution is increased. An ideal solution is defined as one in which the colligative properties vary linearly with solute concentration at all concentrations. If sufficiently diluted, all solutions show this linear variation and can be considered to act ideally. If a nonvolatile solute, such as sodium chloride or glycine, is added to a volatile solvent, such as water, the vapor pressure of the solvent is decreased. The linear relationship of this decrease to the mole fraction of the solvent in solution was formulated by Raoult and is *Raoult's law*:

$$P_A = X_A P_A^0 \tag{13.18}$$

where P_A is the vapor pressure of the solvent, X_A is the mole fraction of the solvent, and P_A^0 is the vapor pressure of pure solvent. If a solution obeys Raoult's law, it is defined as *ideal*.

All of the colligative properties are interrelated and consequently information about the osmotic pressure, vapor pressure, freezing point, and boiling point can be determined from any of the other properties. The relationship of the colligative properties to solute concentration is summarized in the following equations. The freezing point depression for dilute solutions can be expressed as

$$\Delta T_\mathrm{f} = k_\mathrm{f} m \tag{13.19}$$

where k_f is

$$k_\mathrm{f} = \frac{M_\mathrm{A} R T_\mathrm{o}^2}{1000 \Delta H_\mathrm{fusion}} \tag{13.20}$$

where m is the molality of the solute and M_A is the molecular weight of the solvent. For water, $k_\mathrm{f} = 186$. Equation (13.19) is derived from a simplification of the *virial equation*.

The boiling point elevation is given by

$$\Delta T_\mathrm{b} = k_\mathrm{b} m \tag{13.21}$$

with k_b given by

$$k_\mathrm{b} = \frac{M_\mathrm{A} R T_\mathrm{o}^2}{1000 \Delta H_\mathrm{vaporization}} \tag{13.22}$$

For water, $k_\mathrm{b} = 0.51$.

For a volatile solute, the vapor pressure of the solute fraction is given by *Henry's law*:

$$P_\mathrm{B} = k_\mathrm{B} X_\mathrm{B} \tag{13.23}$$

where k_B is the *Henry's law constant* (representative values are listed in Table 13.1).

The osmotic pressure plays an important role in the behavior of cells and in their physiology. Osmotic pressure is defined as the external pressure that must be exerted on a solution to maintain the activity of the solvent at the same value as that of pure solvent (at constant temperature). Consider a system that is comprised of a solution of glucose in water separated by a semipermeable membrane (permeable to water

Table 13.1 Values of k_B for Henry's law

Compound	k_B ($\times 10^3$) (in atmospheres at 298 K)
N_2	86.0
CO_2	1.6
O_2	43.0
CH_4	41.0

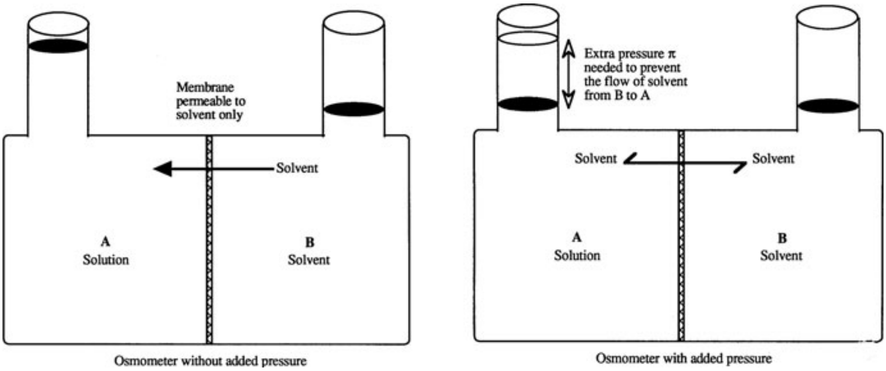

Membrane permeable to solvent only
Solvent
A Solution
B Solvent
Osmometer without added pressure
Extra pressure π needed to prevent the flow of solvent from B to A
Solvent
Solvent
A Solution
B Solvent
Osmometer with added pressure

Fig. 13.2 Diagram of an instrument that measures osmotic pressure, called an osmometer. An osmometer consists of two chambers separated by a membrane permeable to the solvent, but not the solute. A solute added to compartment. A lowers the activity of the solvent in compartment. A compared to that in compartment B. Solvent will flow from B to A (osmosis) unless the activity of the solvent in A can be restored to its previously higher level. Measuring the amount of pressure, π that must be applied to compartment A to raise the activity of the solvent in A to the same level as in B gives the osmotic pressure

but not to glucose) from a quantity of pure solvent (water in this case) (Fig. 13.2). As already described, the addition of solute to pure solvent lowers the activity of the solvent as reflected in alterations in its colligative properties. Therefore, the activity of the water component in the glucose solution is lower than that of the water on the pure solvent side of the apparatus. At equilibrium, the chemical potential of each of the components must be equal, and since only water is free to move across the membrane, water must move from the pure solvent compartment into the solution compartment. The movement of the solvent across the membrane is called *osmosis*. In order to prevent this flow of solvent, the activity of the water on the solution side must be raised so that the equivalence of the chemical potential is restored. The activity of the water in the solution phase can be raised by exerting pressure on the solution side. This is the osmotic pressure, π.

For solutions that are dilute, the *van't Hoff equation for osmotic pressure* may be used to find π:

$$\pi = \frac{nRT}{V} = cRT \tag{13.24}$$

where c is the molar concentration, equal to n/V. The van't Hoff equation is derived in Appendix M. Readers should observe the similarity between the van't Hoff relation and the ideal gas equation and a pleasing symmetry in thermodynamics awaits those who do. If the solute molecules and the gas molecules are considered analogously, and the solvent and the vacuum of the container in which these molecules are free to move are equated, it is easy to see why both systems can be treated equivalently. Activity and fugacity in these ideal systems are equivalent to concentration and pressure, respectively. When there are interactions between the solute molecules or between the solute and solvent, the activities of the solvent and the solute change. A key point about osmotic pressure must be emphasized: the osmotic pressure must not be considered as a pressure exerted by the solute; rather it is the external pressure required to equalize the chemical potential of the solvent, *because it is the solvent to which the membrane is permeable and that will move to equalize the chemical potential*. The process by which the solvent moves through the membrane is not important in finding the osmotic pressure, since the equilibrium result is the same regardless of the mechanism.

All of the colligative properties can be used in the laboratory to find molecular weights and activities of solutions. Osmotic pressure measurements are well suited to measuring the activity of solvents and of molecular weights of macromolecules because of the extremely large magnitude of the osmotic pressure. For example, the osmotic pressure associated with a 1-M solution at 298 K is 24.2 atm, which corresponds to a mercury column of 18.4 m. Molecular weights can be found using osmotic pressure measurements with the following formula:

$$\mathrm{MW}_{(x)} = \frac{RTW_{(x)\,(\mathrm{grams})}}{\pi} \tag{13.25}$$

where again the virial equation is simplified and W is the weight in grams of the substance x being evaluated.

13.4.2 The Activity Coefficient Can Be Measured by Changes in the Colligative Properties

All of the colligative properties are related to one another because of their shared common parentage in Eq. (13.1); the chemical potential of a component in different phases is always equal at equilibrium. The activity of the solvent can be calculated from the colligative properties of a solvent. The most useful of these are vapor pressure and osmotic pressure measurements since they can be done at any temperature. The activity of a solvent is directly available from osmotic pressure measurements:

$$\ln a = \frac{-\pi \overline{V}_{\mathrm{A}}}{RT} \tag{13.26}$$

This formulation resulted as part of the derivation of Eq. (13.24) and is shown in Appendix M. (Derivation of a formula to find activity from vapor pressure measurements is left as an exercise for the reader.)

13.5 Surface Phenomena Are an Important Example of Phase Interaction

When different phases come in contact with each other, an interface between them occurs. This interface is a surface and the properties of a surface are different from the properties of either of the phases responsible for creating it. The region that includes the surface and the immediately adjacent anisotropic boundary is called the *interphase*. Some of the general thermodynamic properties of surfaces will be the focus of this section and the associated interphases will be discussed in Chapter 20.

Why are surfaces any different from the bulk phase from which they are derived? In one sense, they differ because they define where the phase begins or ends. Surfaces are boundaries and boundaries are anisotropic. They are different from the bulk phase simply because they have an inside that is bulk phase and an outside that is not. The molecules making up a surface are different from those of the bulk phase because they are not surrounded by bulk phase molecules. Consider a bulk phase of water molecules. These water molecules may be visualized as being arranged in a tetrahedral array. The cohesive energy from the association of the molecules leads to an energy of association per mole, U. Each water molecule therefore has an energy associated with it:

$$\frac{U}{N_\mathrm{A}} = c \tag{13.27}$$

where N_A is Avogadro’s number. In the tetrahedral array, each molecule is bound to four others and the bond energy is $c/4$. Consider that the water molecules at the phase boundary or surface will have three rather than four neighbors. The bonding energy of the surface molecules will be $3(c/4)$ rather than $4(c/4)$. The bonding energy of the surface molecule is therefore less than the bonding energy associated with a bulk phase molecule and the energy of the surface molecule is therefore higher than that of molecules in the bulk phase. It follows that work must be done to move a molecule from the interior bulk phase of the water to the surface. Because the presence of a surface requires the expenditure of work and a higher energy, all materials attempt to minimize the surface area. The force in the surface that attempts to keep the surface area at a minimum is called the *surface tension*. As a surface is stretched, the free energy of the surface increases and the increase in free energy (usually given in erg/cm^2) is related to the change in area through the surface tension (designated by the symbol g and usually given in dyn/cm). The more highly associated a substance is, the higher the surface tension will be. The surface tensions for various materials are compared with that of water in Table 13.2.

Table 13.2 Surface tensions of liquids (in air) at 298 K

Compound	γ (in mN/m = dyn/cm)
Hg	485.48
H_2O at 298 K	71.99
H_2O at 373 K	58.91
Benzene	28.22
Acetone	23.7
n-Hexane	18.4
Acetaldehyde	20.50
Acetic Acid	27.10
Ethanol	21.97
DMSO	42.92
Cyclohexane	24.65

What happens to the surface tension of a pure solvent such as water when a solute is added? The answer depends on the solute. If the substance decreases the free energy of the surface, then the solute will concentrate at the surface, and the surface tension will decrease. If, on the other hand, the solute were to raise the surface tension of the solution, it would be energetically unfavorable for it to concentrate at the surface, and it will instead shun the surface, preferring the bulk phase. Therefore, substances that are capable of lowering the surface tension will do so quite dramatically, because it is energetically favorable to do so. Material that would be able to raise the surface tension will at best only slightly raise the surface tension, because it is energetically unfavorable for it to be found at the surface. Generally the addition of solute results in either a lowering or a slight increase in surface tension.

The concentration of solute molecules at a surface is different than in the bulk. The surface excess concentration, denoted by the symbol Γ, represents the quantity of solute adsorbed to the surface. The surface excess can be quantified through the *Gibbs adsorption isotherm*:

$$\Gamma_i = \frac{-1}{RT}\left(\frac{\partial \gamma}{\partial \ln a_i}\right) \tag{13.28}$$

Γ has the units of mol/m^2. The surface tension is γ and the activity of the solute, *i*, in solution is a_i. This expression indicates that a solute that decreases the surface tension will concentrate at the surface, since the sign of the excess concentration is opposite that of the change in surface tension.

Some of the most important surface-active compounds in biological systems are the amphiphilic long-chain fatty acids. By virtue of their hydrophilic–COOH and hydrophobic hydrocarbon groups, these molecules will preferentially adsorb to the water surface with the hydrophobic tail sticking out of the water phase. They act to lower the surface tension of the aqueous phase significantly and are important, for example, in the lungs where they increase the surface area available for gas exchange. These amphiphilic fatty acids or detergents will form surface films of monomolecular dimension if allowed to spread on the surface of water. The change in surface tension that results from the formation of these monolayers can be easily measured through the use of a Langmuir film balance or similar devices. Multilayers of these detergent molecules can be formed by the repeated dipping of a glass slide through a monolayer. As the slide is repeatedly passed through the monolayer, layer after single layer of molecules is deposited, one upon the other, polar to polar and nonpolar to nonpolar portions of the molecule orienting toward each other. The resultant film is called a Langmuir–Blodgett film, and its thickness can be directly measured by optical interference methods. Since the number of dips is related to the number of monolayers, the length of the detergent molecule can be directly calculated. Biological membranes are generally considered as being comprised of a bilayer of amphiphilic phospholipid molecules. These phospholipids are composed of a glycerol backbone, nonpolar hydrocarbon chains, and a polar head group, and thus resemble a Langmuir–Blodgett film generated by two passes through a monolayer. In the biological membrane, there are a variety of different lipid components, as well as proteins and polysaccharides, making up the membrane and giving it a wide range of properties. The formation and properties of the biological membrane will be taken up in more detail in Chapter 17.

The change in surface tension is reflective of a change in surface free energy and consequently many other properties of the surface are changed as well. Extremely important biological effects are seen in the varied transport of small molecules, such as water and ions, and in the fluidity of the membrane, which affect the lateral diffusion of large molecules such as proteins that act as receptors. Membranes can have compositions that result in physical properties similar to those found in liquids or in solids and can undergo phase transitions. Since change in surface tension is a direct measure of free energy change, the thermodynamics of the phase changes can be determined through the measurement of the surface tension under varying conditions. In later sections a more detailed picture of the role of surfaces and membranes will be developed with regard to their role in transport, information transduction, mechanical work functions, and modulation of cellular function. In addition, the properties and forces acting in the interphase region, a very important area that is associated with surfaces, will be examined. The properties of the interphase are fundamental to the accurate description of the biochemistry and the biophysics of cellular behavior, because the volume of the cell represented by the interphase region is enormous. This is the result of the simple fact that cells have an enormous surface area attributable to the large number of cellular membranes.

13.6 Binding Equilibria Relate Small Molecule Binding to Larger Arrays

A very important phenomenon in biological systems is the *binding* of small molecules to a larger molecular array, usually a macromolecule or a surface. Binding processes are viewed as a special type of equilibrium experiment. Most cases of macromolecular binding are characterized by

(1) multiple binding sites for the ligand on the macromolecule. Therefore a mole of the macromolecule *P*, will bind a number of moles of the ligand *A*, at a number of sites, *n*.
(2) characteristic equilibrium binding constants for the interaction between the sites and the ligand. In many cases there will be classes of binding sites n_j, that are thermodynamically distinguishable, contained in the macromolecule. The properties of the binding sites may be independent or dependent on the prior binding of ligand to other binding sites.
(3) dependence on the binding of a different ligand to another binding site elsewhere on the macromolecule.

Equilibrium binding experiments explore these characteristics using either the method of equilibrium dialysis (vide infra) or a physical method that is sensitive to the change in the character of either the ligand or the macromolecule upon binding. These methods often use spectroscopic methods such as absorption and fluorescence to probe the interaction. In the case of equilibrium binding we are interested in the total concentration of ligand which is equal to the sum of the free and bound ligands:

$$[A_\mathrm{t}] = [A_\mathrm{f}] + [A_\mathrm{b}] \tag{13.29}$$

We can define the average number of ligands bound to the macromolecule *v*, in relation to the molar concentration of the macromolecule and the concentration of the bound ligand:

$$v = \frac{[A_\mathrm{b}]}{[P_\mathrm{t}]} \tag{13.30}$$

A graph of [A] against *v* will, in general, reach a maximum at some finite number *n* (Fig. 13.3). The analysis of this type of experiment is usually made useful by special graphical transformations such as the Scatchard plot.

In the second case where a physical parameter *X* is measured to report on the binding action, it must be assumed or shown that the change in this physical parameter is linear with respect to the binding and is the same at each binding site. As we will learn when we discuss of spectroscopy (Chapter 28), these techniques are very sensitive to the environment and so these assumptions can be somewhat tenuous if spectral analysis is used as the physical probe. However, if these properties can be demonstrated, the ratio of the change in *X* with partial binding with respect to *X* when all sites are bound will give a quantity called the *fraction saturation*:

$$\frac{\Delta X}{\Delta X_\mathrm{T}} = \frac{v}{n} = \theta \tag{13.31}$$

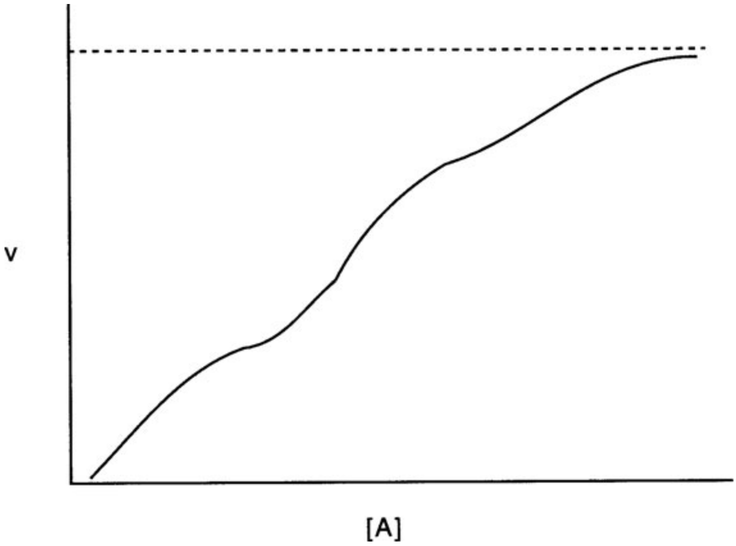


Fig. 13.3 A graph of [*A*] against *v* will, in general, reach a maximum at some finite number *n*

13.6.1 Binding at a Single Site

The simplest case is the binding of a single ligand per macromolecule. This case was essentially treated in our earlier discussion of the equilibrium constant. To review briefly we are interested in the relationship:

$$P + A \rightleftarrows PA \tag{13.32}$$

where either a binding constant or a dissociation constant can be written as

$$K_{\text{assoc}} = \frac{[PA]}{[P]\,[A]} \qquad K_{\text{dissoc}} = \frac{[P]\,[A]}{[PA]} \tag{13.33}$$

For the free energy dependence we write

$$\Delta G^{\text{o}} = -RT \ln K_{\text{assoc}} \tag{13.34}$$

We can relate these equations to the observable *v* for this special case of a single binding site (which is therefore independent):

$$v = \frac{[A]\,/K_{\text{assoc}}}{1 + [A]\,/K_{\text{assoc}}} \tag{13.35}$$

The derivation of this equation can be found in the next section, where we will consider the more general case of ligand binding to multiple independent sites.

13.6.2 Multiple Binding Sites

We know that the more typical case of macromolecular binding involves multiple sites. While *v* represents the average number of sites per macromolecule, each macromolecule will contain some number of ligands between 0 and *n*. Thus the total concentration of bound ligand will be

$$\begin{aligned}[A_\mathrm{b}] &= [PA] + 2\,[PA_2] + 3\,[PA_3] + \ldots n\,[PA_n] \\ &= \sum_{i=1}^{n} i\,[PA_i]\end{aligned} \tag{13.36}$$

The total molar concentration of the macromolecule includes the unbound state:

$$[P_\mathrm{t}] = \sum_{i=0}^{n} i\,[PA_i] \tag{13.37}$$

The general case for *v* is then

$$v = \frac{[A_\mathrm{b}]}{[P_\mathrm{t}]} = \frac{\sum\limits_{i=1}^{n} i\,[PA_i]}{\sum\limits_{i=0}^{n} i\,[PA_i]} \tag{13.38}$$

We must write a series of equilibrium relationships between the macromolecule and its various liganded states [*PAi*]. This is just a series of equations similar to Eq. (13.32):

$$\begin{array}{ll} P + A \leftrightarrow PA & K_1 = \dfrac{[PA]}{[P]\,[A]} \\ P + 2A \leftrightarrow PA_2 & K_2 = \dfrac{[PA_2]}{[P]\,[A]^2} \\ P + 3A \leftrightarrow PA_3 & K_3 = \dfrac{[PA_3]}{[P]\,[A]^3} \\ \cdot & \\ \cdot & \\ \cdot & \\ P + nA \leftrightarrow PA_n & K_n = \dfrac{[PA_n]}{[P]\,[A]^n} \end{array} \tag{13.39}$$

This allows us to rewrite Eq. (13.38) in terms of the equilibrium constants:

$$v = \frac{\sum\limits_{i=1}^{n} K_i\,[P]\,[A]^i}{\sum\limits_{i=0}^{n} K_i\,[P]\,[A]^i} \tag{13.40}$$

K_0 is unity and $[P]$ factors out:

$$v = \frac{\sum\limits_{i=1}^{n} iK_i\,[A]^i}{\sum\limits_{i=0}^{n} K\,[A]^i} \tag{13.41}$$

This general equation is known as the Adair equation. We will use it as a starting point to consider several specific cases: (1) binding when all sites are equal and independent, (2) binding when sites are non-equivalent, and (3) binding when the sites influence one another (cooperativity).

13.6.3 Binding When Sites Are Equivalent and Independent

This is the simplest case of multiple equilibrium. A macromolecule has a number of sites to which a ligand may bind. For now we assume the affinities of each site are independent of each other. We wish to quantitate the number of each class of liganded macromolecule (some with one ligand, some with two, etc.). Since there are a variety of possible arrangements of ligands complexed to a macromolecule we must treat this problem as a statistical distribution Figure 13.4 shows the possible ways that two ligands could be bound to a macromolecule with four binding sites. There are six isomeric forms in this example and in general the number of isomers can be determined from

$$N_{i,n} = \frac{n!}{(n-1)!i!} \tag{13.42}$$

The existence of these isomeric forms is not recognized by the macroscopic equilibrium expression that we have written above. However, the expression for the affinity of a ligand for a particular site must take the specific class of isomer into account. We want to be able to write a microscopic equilibrium expression for specific path: the addition of a ligand to a particular isomer k, which is a member of the class PA_{i-1}, which with binding becomes the isomer l of the class PA_i:

$$(k_i)_{kl} = \frac{\left[PA_{i,l}\right]}{\left[PA_{i-1,k}\right][A]} \tag{13.43}$$

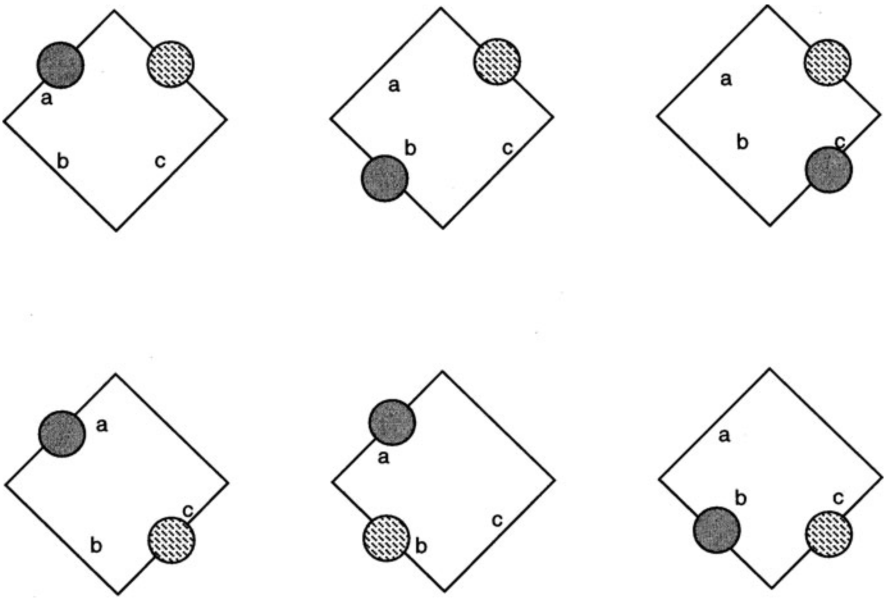


Fig. 13.4 Two ligands can be bound in six conformations to a macromolecule with four binding sites

The various ways that the microscopic addition can be made depend on *i*, and so the sums written in Eq. (13.41) will be unequal even in the special case chosen here where all of the microscopic binding constants are the same. We must sum over all of the possible paths given by Eq. (13.42). Since all of the microscopic binding constants are the same we will write them all in terms of *k*:

$$[PA_i]_l = k^i[P][A]^i \tag{13.44}$$

All of the isomeric forms will have the same average concentration so we multiply times the distribution factor (Eq. 13.42):

$$[PA_i] = \frac{n!}{(n-1)!i!}k^i\,[P]\,[A]^i \tag{13.45}$$

v can be written in terms of this expression for the total concentration of PA_i:

$$v = \frac{\sum\limits_{i=1}^{n} \dfrac{n!}{(n-1)!i!}k^i\,[A]^i}{\sum\limits_{i=0}^{n} \dfrac{n!}{(n-1)!i!}k^i\,[A]^i} \tag{13.46}$$

This expression simplifies by recognizing the denominator as the binomial expansion of $(1 + k[A])^n$. The numerator simplifies as well with the final result as

$$v = \frac{nk\,[A]}{1 + k\,[A]} \tag{13.47}$$

Thus as in the expression for the case of the single binding site, the binding of a ligand to each of several independent sites will yield a straight line when converted to the form of the Scatchard equation. We will now explore how this analysis can be used in an equilibrium dialysis experiment.

13.6.4 Equilibrium Dialysis and Scatchard Plots

In an equilibrium dialysis experiment, the binding of a small molecule, *i*, to a macromolecule, *M*, is studied by establishing two phases through the use of a dialysis membrane. The macromolecular species is placed inside the dialysis bag and it is suspended in a large volume (usually >100 times the volume in the dialysis bag). There are therefore two phases, phase *a* which is *inside* the bag and phase *b* which is *outside* the bag. The concentration of the small molecule is the measured quantity, C_i. At equilibrium, the chemical potential of *i* is the same in both phases:

$$\mu_{i(a)} = \mu_{i(b)} \tag{13.48}$$

The activity of *i* in each phase can also be equated:

$$a_{i(a)} = a_{i(b)} \tag{13.49}$$

Since the activity is equal to the product of the activity coefficient, γ, and the concentration of *i*, c_{i}, the following relation can be written:

$$\gamma_i^a c_{i(a)} = \gamma_i^b c_{i(b)} \tag{13.50}$$

The solution in phase *b* is made dilute enough so that the activity coefficient γ_i^b can be considered equal to unity; therefore, Eq. (13.50) can be simplified and rearranged as

$$\gamma_{i(a)} = \frac{c_{i(b)}}{c_{i(a)}} \tag{13.51}$$

The activity coefficient of *i* inside the bag is generally less than unity, indicating that there is a greater concentration of *i* in phase *a* than in phase *b*. The interpretation of this thermodynamic measurement is that the difference in the amount of *i* is due to the binding of *i* to the macromolecule. If the activity coefficient of *i* that is free inside the bag is assumed to be unity, then the concentration of $i_{\text{free inside}}$ is equal to the concentration of i_{outside}:

$$c_{i\text{-free}(a)} = c_{i(b)} \tag{13.52}$$

but

$$c_{\text{inside}(a)} = c_{\text{bound}(a)} + c_{\text{free}(a)} \tag{13.53}$$

Rearranging Eqs. (13.52) and (13.53) gives the result:

$$c_{i\text{-bound}(a)} = c_{i\text{-inside}(a)} - c_{i(b)} \tag{13.54}$$

Measuring the total concentration of *i* inside and outside and subtracting the difference gives the amount of *i* bound. The amount bound, *v*, per macromolecule, *M*, can then be expressed as

$$v = \frac{c_{i\text{-bound}(a)}}{c_{\text{M}}} \tag{13.55}$$

where *v* is equal to the average number of molecules of *i* bound to each macromolecule.

v will depend on the binding constant at equilibrium, the number of binding sites on a macromolecule, and the actual concentrations of *M* and *i*. If the total number of sites available for binding is given by *N*, then θ, defined earlier as the fraction of sites bound is

$$\theta = \frac{v}{N} \tag{13.56}$$

This analysis will assume that the binding sites are independent and identical to one another and so the relationship derived for a single binding site is the same for all binding sites. Ignoring activity coefficients, at equilibrium, a binding constant *K* can be written for the reaction:

$$M + i \leftrightarrow M - i \tag{13.57}$$

giving

$$K = \frac{[M - i]}{[M]\,[i]} \tag{13.58}$$

Equation (13.56) could be rewritten in terms of the concentrations of *M* and *i*:

$$\theta = \frac{[M - i]}{[M] + [M - i]} \tag{13.59}$$

Now, Eqs. (13.58) and (13.59) can be combined to give the fraction of sites bound in terms of the equilibrium binding constant, *K*, for all *N* when binding sites are equivalent:

$$\theta = \frac{K\,[i]}{1 + K\,[i]} \tag{13.60}$$

The binding constant of [*i*] is given by

$$\frac{\theta}{1-\theta} = K\,[i] \tag{13.61}$$

Since the measured quantities in an equilibrium binding experiment are the concentrations of *i* inside and outside the dialysis bag, and *N* and *K* are generally the coefficients being sought, Eq. (13.61) is rewritten in terms of *v* and *N*:

$$\frac{v}{N-v} = K\,[i] \tag{13.62}$$

The *Scatchard equation* is obtained by rearranging into a $y = mx + b$ form:

$$\frac{v}{[i]} = K(N-v) \tag{13.63}$$

Plotting ${}^{v}\!/_{[i]}$ on the *y*-axis against *v* on the *x*-axis (the numbers are easily available from experiment) will give a straight line, at least at low values of *v* (Fig. 13.5). The *x*-intercept of this plot is *KN*, the number of binding sites per molecule, and the slope of the line is –*K*. A nonlinear plot of Eq. (13.63) is evidence that the binding sites on the macromolecule under study are not identical, equivalent, or independent; that is, they exhibit cooperativity.

A final point to be made about the special case of the multiple independent ligand binding is that Eq. (13.63) can be written as

$$\frac{{}^{v}\!/_{N}}{1 - {}^{v}\!/_{N}} = K[A] \tag{13.64}$$

${}^{v}\!/_{N}$ is the fraction saturation which we have designated as θ. Thus we can obtain from Eq. (13.65) an equation of the form of (13.61):

$$\frac{\theta}{1-\theta} = K\,[A] \tag{13.65}$$

This equation is the *Langmuir adsorption isotherm* and describes the thermodynamics of surface coverage or binding in the case of non-interacting binding sites on a surface. Thus another close parallel between surface chemistry and macromolecular behavior can be seen.

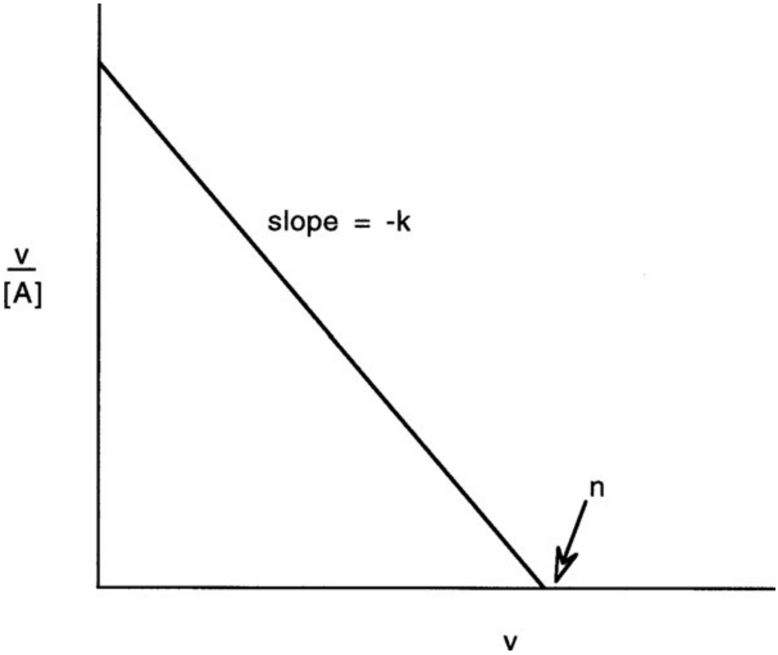


Fig. 13.5 Idealized Scatchard plot analysis. The *x*-intercept labeled n in the figure is equivalent to NA as described in the text

13.6.5 Binding in the Case of Non-equivalent Sites

We have treated the microscopic binding constants as equivalent. Now we consider the case in which the binding sites are different, some weak and some strong. We will still consider them as independent. There are N classes of site numbered as $S = 1 \rightarrow N$. Within each class there are n_s sites per class. v will be the sum of each of each v_s value:

$$v = \sum_{s=1}^{N} v_s \tag{13.66}$$

In other words the value of v will be the sum of the numbers of each class of site occupied per macromolecule. Each class is independent with a characteristic k_s value, so in terms of Eq. (13.47) we write

$$v_s = \frac{n_s k_s [A]}{1 + k_s [A]} \tag{13.67}$$

The overall v is the sum of Eq. (13.47):

$$v = \sum_{s=1}^{N} \frac{n_s k_s [A]}{1 + k_s [A]} \tag{13.68}$$

We consider a system with two classes of binding sites, one strong and one weak. There are n_1 sites with strong binding constant k_1 and n_2 sites with weak constant k_2. k_1 is greater than k_2. Equation (13.68) is

$$v = \frac{n_1 k_1 [A]}{1 + k_1 [A]} + \frac{n_2 k_2 [A]}{1 + k_2 [A]} \tag{13.69}$$

This equation is a weighted mixture of two lines with particular slopes. Unless k_1 is much greater than k_2 there will be a fairly smooth mixing of their characters and it will be hard to tell if there are actually two populations or just one population with intermediate k. In the extreme cases the distinct slopes of each population can be seen (Fig. 13.6a). The slopes are much more easily appreciated with a Scatchard plot (Fig. 13.6b).

Another useful graphical rendering is the *Hill plot* in which log $\frac{\theta}{1-\theta}$ is plotted against log $[A]$ (Fig. 13.7). The Hill plot causes the lowest values of $[A]$ to fall near a line that represents the behavior of the strong binding sites. As $[A]$ increases, the data points move to the vicinity of a line that represents the behavior of the weak binding sites. The Hill plot cannot provide information about the number of binding sites, but it does allow the equilibrium constants to be estimated and shows the behavior of the overall system.

13.6.6 Cooperativity Is a Measure of Non-independent Binding

If the binding at one site increases the likelihood of binding at the next site, a positive *cooperativity* exists. If in contrast the affinity for the subsequent binding event falls, a negative cooperativity exists. This phenomenon is called *allostery* and includes effects both on binding sites of the same class (*homeoallostery*) and of different classes (*heteroallostery*). While the plot of non-cooperative binding is a curve of a rectangular hyperbola, cooperative binding has a characteristic sigmoid shape. The sigmoid shape results from the conversion of sites with relatively low affinities into sites with much higher affinities and hence a steeper slope. These effects can be seen dramatically on a Hill plot in which the data points move from the low affinity sector to the region of higher affinity with increasing $[A]$ (Fig. 13.8). A Hill plot showing negative cooperativity looks like the case of the high- and low-affinity independent sites (Fig. 13.7). The slope and shape of the curve that results from a Hill plot are useful in determining the type of binding seen (Fig. 13.8). We have seen already that a curve with a positive slope near the midpoint indicates positive cooperativity and that a curve with a negative slope indicates either negative cooperativity or a system with greater than one class of sites. A straight line plot with a slope of 1 is diagnostic

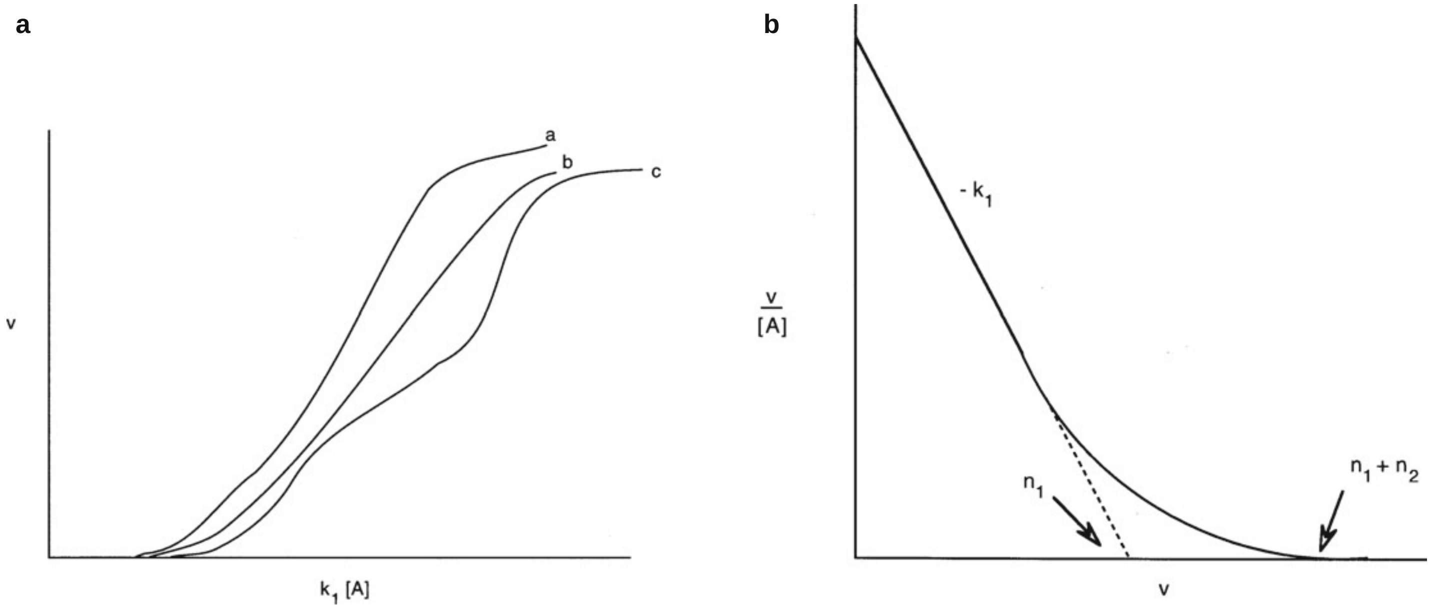


Fig. 13.6 (**a**) Binding curve for a macromolecule with two classes of non-equivalent binding sites. The behavior of the separate classes can be hard to see clearly on a graph of this type and (**b**) often reexpressing the data on a Scatchard plot helps in the analysis

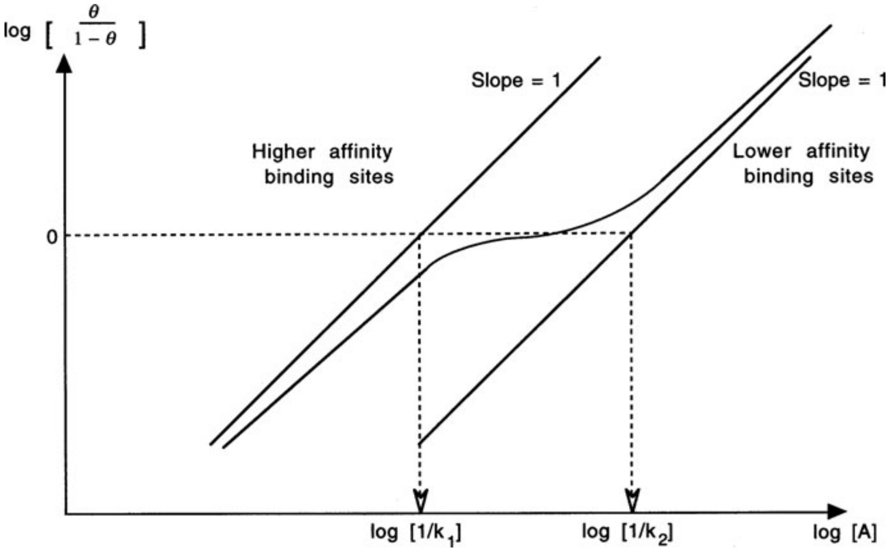


Fig. 13.7 The Hill plot. Shown is the case for a macromolecule with both high and low affinity sites of independent nature

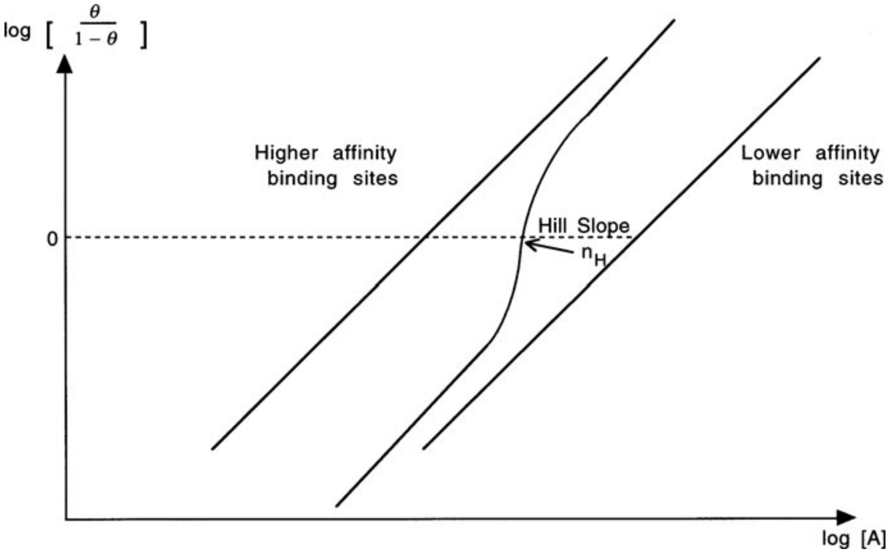


Fig. 13.8 Hill plots can be used to ascertain the cooperative properties of a system

of a system with independent and equivalent binding sites. The slope of a Hill plot has a maximum of *n*, the number of sites on the macromolecule. The slope actually found, n_H, is always less than *n*. The more closely n_H approaches *n* the higher the cooperativity of the interaction.

There are two dominant models that have been proposed to describe the phenomenology of cooperativity. The first of these supposes that changes in subsequent ligand binding occur in all the binding sites in concert. Thus there is a smooth transition from the weak affinity state, *T* into the strong affinity state, *R*. This model was proposed in 1965 by Monod, Wyman, and Changeux and is named the MWC model. The second model supposes that the changes occur sequentially with the result that a mixed state of weak affinity sites *A* and strong affinity sites *B* exists. This model was proposed in 1966 by Koshland, Nemethy, and Filmer and is named the KNF model.

The MWC model does not allow any mixture of the *T* and *R* states. A molecule is either in *T* form or in *R* form. The binding constant for the *R* form is larger than for the *T* form and there is a preferred binding to the *R* form. The concerted action occurs because the binding to one low affinity ligand increases the probability that all of the subsequent binding sites will be in the *R* state and hence the affinity of the macromolecule will be significantly increased. Without the presence of ligand, an equilibrium denoted by *L* is said to exist:

$$L = \frac{[T]}{[R]} \tag{13.70}$$

L is assumed to be a large number hence the dominant species will be the *T* form. The ratio of the binding constants for the two states is given by *c*, which is thought to be a small number:

$$c = \frac{k_T}{k_R} \tag{13.71}$$

When the ligand binds to the macromolecule, *L* changes by a factor *c*. When two ligands bind, the factor changes by c^2. In general the ratio will change

$$\frac{[T_i]}{[R_i]} = c^i L \tag{13.72}$$

We can write a series of equations similar to Eq. (13.46) to get v in terms of the distributions of *T* and *R*. After some algebra we can write v, the fraction of sites containing ligand in terms of the ligand concentration $[A]$. n is the number of sites and $\alpha = [A]/k_r$.

$$v = \frac{\alpha\,(1+\alpha)^{n-1} + Lc\alpha\,(1+ca)^{n-1}}{(1+\alpha)^n + L\,(1+c\alpha)^n} \tag{13.73}$$

The fraction of molecules in the *R* form is given by

$$\frac{(1+\alpha)^n}{(1+\alpha)^n + L\,(1+c\alpha)^n} \tag{13.74}$$

The limiting behavior of this equation can be seen if we consider what happens when $[A] \rightarrow 0$ and $[A] \rightarrow \infty$. If L is large, then as $[A] \rightarrow 0$ the T state is favored because α is small and the equilibrium constant of T dominates the behavior; the molecule has low affinity for the ligand. If c is small, then as $[A]$ increases, the equilibrium constant of the R form dominates the behavior and the molecule develops a strong affinity for the ligand (Fig. 13.9).

The mathematics of the KNF model is more complicated because the sequential model depends on interactions between pairs of subunits undergoing interaction from A to B state. Subunits are only allowed to be in either state A or B, but overall,

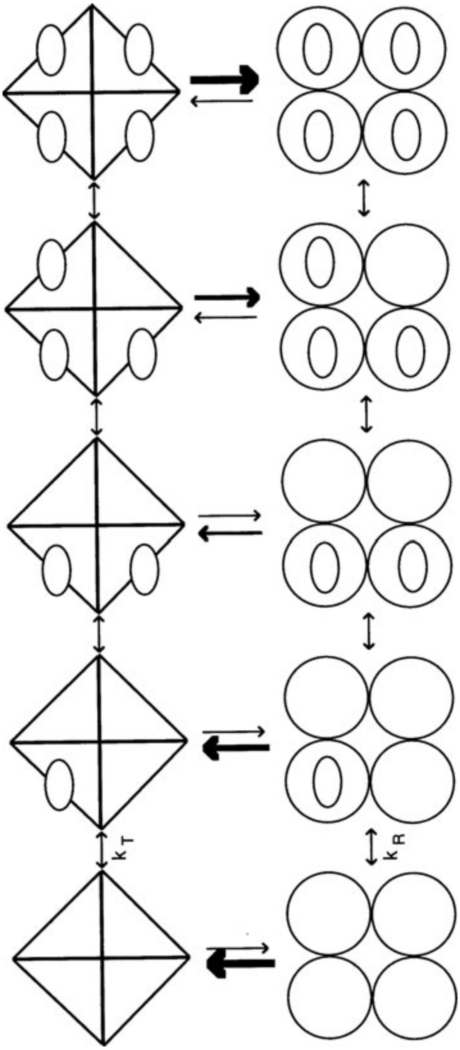


Fig. 13.9 The MWC model of allostery

the macromolecule is allowed to have both *A* and *B* subunits. In contrast to the MWC model where mixed states are forbidden, mixed states are expected in the KNF model. The equilibrium between the different pairs would be variable and a very wide range of control behaviors would be expected. In this model, when a ligand binds to a site it induces that subunit to go from state *A* to *B*. The presence of the *B* subunit near other *A* subunits may then induce those *A* subunits to increase or decrease affinity for sequential binding to their ligands. The pattern of this model is illustrated in Fig. 13.10.

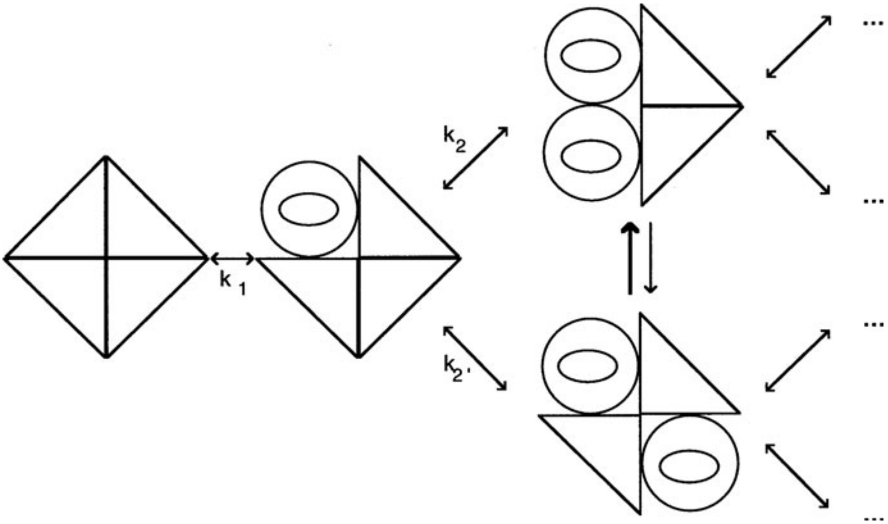


Fig. 13.10 The KNF model of allostery

13.6.7 The Acid–Base Behavior of Biomolecules Reflects Proton Binding

The dissociation of chemical species into an ion and proton is expressed in terms of an equilibrium constant for that reaction. In aqueous solutions the foundation reaction is the ionization of water:

$$H_2O \rightleftharpoons H^+ + OH^- \tag{13.75}$$

The equilibrium constant is K_w and is given by

$$K_w = \frac{\left[H^+\right]\left[OH^-\right]}{[H_2O]} \tag{13.76}$$

The amount of water consumed by ionization is very small compared to the molar concentration of water, so the water concentration is assumed to have an activity of 1, and Eq. (13.76) is written in terms of the ion product, $K_\mathrm{w} = 10^{-14}$. When neither acid nor base predominates, $[\mathrm{H}^+] = [\mathrm{OH}^-] = 10^{-7}$. The pH is 7 because

$$\mathrm{pH} = -\log[\mathrm{H}^+] \tag{13.77}$$

For the dissociation of a weak acid

$$\mathrm{HA} \rightleftharpoons \mathrm{H}^+ + \mathrm{A}^- \tag{13.78}$$

then the equilibrium expression is

$$K_\mathrm{a} = \frac{[\mathrm{H}^+][\mathrm{A}^-]}{[\mathrm{HA}]} \tag{13.79}$$

The pK_a is the pH at which half of the acid (HA) is dissociated. The Henderson–Hasselbalch equation defines the relationship between the pH and the pK_a. We rearrange Eq. (13.79) as

$$\frac{1}{[\mathrm{H}^+]} = \frac{1}{K_\mathrm{a}}\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \tag{13.80}$$

Taking the log of both sides gives

$$\log\frac{1}{[\mathrm{H}^+]} = \log\frac{1}{K_\mathrm{a}} + \log\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \tag{13.81}$$

Substituting gives the Henderson–Hasselbalch equation as

$$\mathrm{pH} = \mathrm{p}K_\mathrm{a} + \log\frac{[\mathrm{A}^-]}{[\mathrm{HA}]} \tag{13.82}$$

A weak acid is only partially dissociated depending on the pH of its environment. A titration experiment will generate the familiar sigmoid-shaped curve in which a buffering region around the pK_a point is found (Fig. 13.11). Most biological acids (and bases) are only partially dissociated and so will exhibit buffering activity near their pK_a points. These values are tabulated in Table 18.7. The titration experiment should be recognized as a binding study and the sigmoid curve should be reminiscent of a Hill plot. We can show that the Henderson–Hasselbalch equation is in equivalent to the Hill equation. In our discussions of binding we focused on association constants while the acid K_a values are dissociation constants. Also we are more interested in the number of dissociated protons r compared to the fully protonated molecules. This is the opposite to our treatment of v above. For a macromolecule

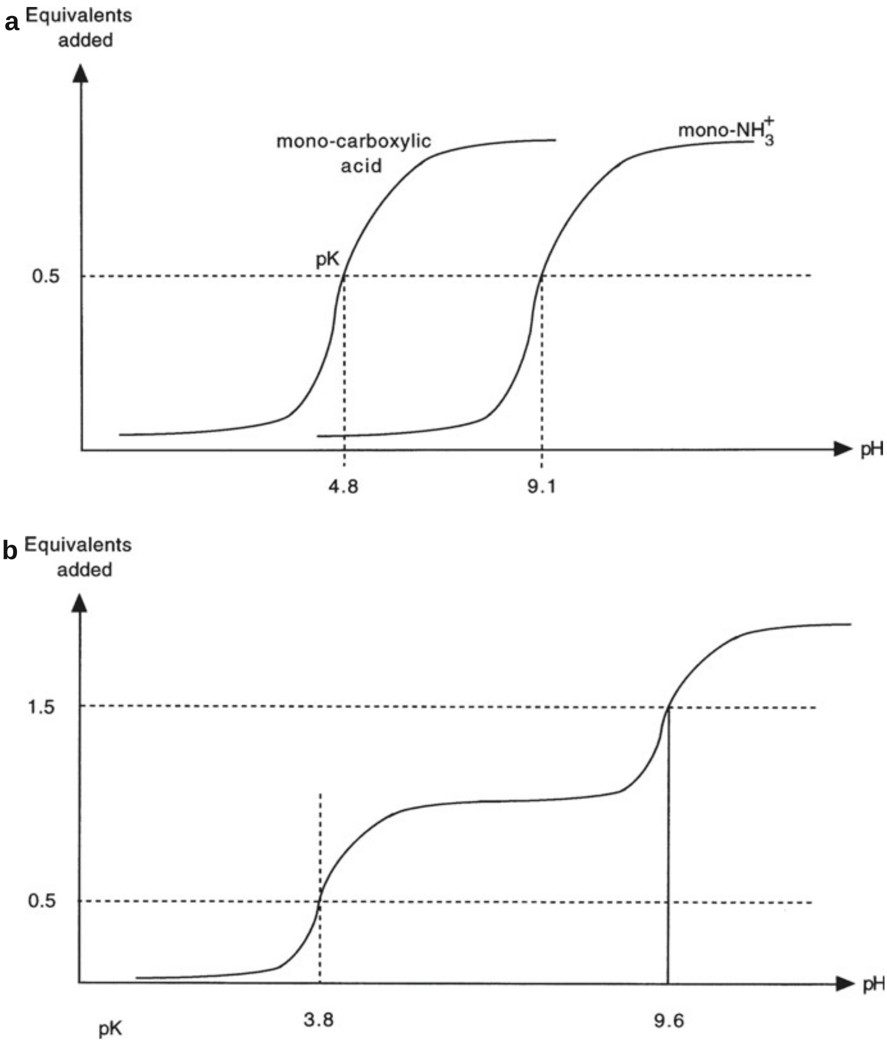


Fig. 13.11 pH binding curve for (**a**) a monoprotic organic acid and a monoprotic organic base. (**b**) an amino acid which is bi-functional and behaves as a binding experiment for two independent binding sites with different affinities

such as a protein with multiple ionizable amino acid side chains *n*, *r* will be *n*−*v*. We consider the case of independent binding sites and using a microscopic dissociation constant *k*, can write

$$v = \frac{n[H^+]/k}{1 + [H^+]/k} \tag{13.83}$$

in terms of r, which is a measure of [H$^+$]:

$$\begin{aligned} r &= n - v \\ &= n\left(1 - \frac{\left[H^{+}\right]/k}{1+\left[H^{+}\right]/k}\right) \end{aligned} \tag{13.84}$$

After some algebra

$$r = \frac{nk/[H^{+}]}{1 + k/[H^{+}]} \tag{13.85}$$

Now r is the fraction of titratable protons dissociated from the macromolecule, so it is similar to the fraction association term θ and we can write in terms of this fraction α:

$$\frac{r}{n-r} = \frac{\alpha}{1-\alpha} = \frac{k}{[\mathrm{H}^{+}]} \tag{13.86}$$

Now taking the logarithm of both sides gives

$$\log\frac{\alpha}{1-\alpha} = \log k - \log\ [\mathrm{H}^{+}] \tag{13.87}$$

which is the same as

$$\log\frac{\alpha}{1-\alpha} = \mathrm{pH} - \mathrm{p}K_{\mathrm{a}} \tag{13.88}$$

The equivalency of the Hill and Henderson–Hasselbalch equations is evident.

In the titration of a polyelectrolyte like a protein we can gather together all of the various classes of binding sites (histidine, lysine, glutamic acid, etc.) and attempt to find the multiple binding equilibria as we did earlier. Such efforts are usually frustrated because of the substantial influence of the local environment on the local pK_a of each site. Furthermore, these dissociations are ionizations and with each proton bound or unbound the charge of the molecule and distribution of the local electrostatic field changes. Thus the free energy change for dissociating each proton from the molecule will be altered by the events preceding it. The overall ΔG term will necessarily include both the intrinsic free energy change associated largely with the isolated group's K_a and also an electrical work term reflecting the extra work needed to leave given the net charge of the macromolecule as

$$\Delta G^{\mathrm{o}} = \Delta G^{\mathrm{o}}_{\mathrm{intrinsic}} + \Delta G^{\mathrm{o}}_{\mathrm{electrical}} \tag{13.89}$$

Using $\Delta G^{\circ} = -RT\ln k$ we write

$$-RT\ln k = -RT\ln k_{\mathrm{in}} + \Delta G^{\mathrm{o}}{}_{\mathrm{elect}} \tag{13.90}$$

which yields

$$\log k = \log k_{\text{in}} - \frac{\Delta G^{\text{o}}_{\text{elect}}}{2.303RT} \tag{13.91}$$

and

$$\text{p}K_{\text{a(apparent)}} = \text{p}K_{\text{a(intrinsic)}} + \frac{\Delta G^{\text{o}}_{\text{elect}}}{2.303RT} \tag{13.92}$$

The electrical term will depend on the electrostatic environment and the net charge z of the macromolecule. If z is positive, the polyelectrolyte will have a positive charge which will enhance dissociation, and so $\Delta G^{\circ}{}_{\text{electric}}$ will be negative.

Further Reading

General

Adamson A.W. (1990) *Physical Chemistry of Surfaces*, 5th edition. Wiley, New York.

Freeman G.R. (ed.) (1987) *Kinetics of Non-homogeneous Processes*. Wiley-Interscience, New York.

Treptow R.S. (1993) Phase diagrams for aqueous systems, *J. Chem. Educ.*, **60**:616–620.

Worley J.D. (1992) Capillary radius and surface tensions, *J. Chem. Educ.*, **69**:678–670.

Binding Interactions

Attle A.D. and Raines R.T. (1995) Analysis of receptor–ligand interactions, *J. Chem. Educ.*, **72:**119–124. (A nice review of the topic including the experimental tools used in biochemical laboratories.)

Koshland, D.L., Nemethy G., and Filmer D. (1966) Comparison of experimental binding data and theoretical models in proteins containing subunits, *Biochemistry*, **5**:365–385.

Monod J., Wyman J., and Changeux J.P. (1965) On the nature of allosteric transitions: a plausible model. *J. Mol. Biol.*, **12**:88–114.

Parody-Morreale A., Cá mara-Artigas A., and Sá nchez-Rulz J.M. (1990) Spectrophotometric determination of the binding constants of succinate and chloride to glutamic oxalacetic transaminase, *J. Chem. Educ.*, **67**:988–990.

Wyman J. and Gill S.J. (1990) Binding and linkage. *Functional Chemistry of Biological Macromolecules*. University Books, South Orange, New Jersey.

Problem Sets

1. The process of breathing depends on the movement of air into and out of the lungs. This is accomplished by the piston-like action of the diaphragm, which leads to an expansion and contraction of the lung space. Thoracic pressure varies with lung volume and air moves in and out due to pressure gradients.

(a) Write a differential equation that applies to the free energy changes associated with this system.

(b) What effect, if any, does respiration have on the transport of O_2 and CO_2 across the alveolar walls. Discuss in terms of the chemical potentials of the molecules in this system.

(c) Write an expression for the free energy of transport of O_2 and CO_2 from the lungs to the red blood cells.

2. Lysis of red cells can occur when distilled water is given intravenously. Assuming the internal concentration of red cells is 0.15 M NaCl and the membrane is permeable only to water, what is the reason for the lysis? Quantitate how much force is acting to rupture the cells.
3. The interior of the nucleus and the cytoplasm of a cell can be treated as a pair of phases separated by a membrane. Derive and write an expression for the free energy change associated with the transport of a strand of messenger RNA from the nucleus to the cytoplasm.
4. Derive a formula that allows the determination of the activity of a solvent from the vapor pressure of the solvent.
5. Why can the assumption of an activity coefficient of unity in the phase outside the dialysis bag be made in equilibrium binding experiments? What would be the effect on a Scatchard analysis if this assumption were incorrect?
6. You are given a white powder and are informed that it will form a nonionic aqueous solution. You weigh it and dissolve 30 g in 1 l of water. The solution is placed in an osmometer and an osmotic pressure of 55 atm is measured.

(a) What is the molecular weight of the powder?

(b) The solution is heated to boiling. What is the measured temperature at which it boils?

(c) 30 g of a second white powder is dissolved in another liter of water. When this solution is placed in the osmometer an osmotic pressure of 11 atm is measured. What conclusions may be drawn about the second powder?

(d) If each solution is arranged so that they are separated by a semipermeable membrane, which way through the membrane will solvent flow?

7. An important measure of poisonous substances in the blood is the serum osmolarity.

(a) Based on a normal serum composition as indicated below, what is the expected serum osmolarity?

(b) The legal limit for blood alcohol that defines "being under the influence of alcohol" for ethanol is > 100 mg per 100 ml of serum. Death may occur when the ethanol level rises above 500 mg/100 ml. What is the serum osmolarity at these levels?

(c) Some desperate alcoholics will drink anti-freeze if they cannot get a drink. Anti-freeze contains ethylene glycol, a sweet-tasting polyalcohol that is metabolized to oxalic acid. Oxalic acid can cause kidney failure and severe metabolic acidemia leading to circulatory collapse thus rapid diagnosis needs to be made when a patient appears with a high serum osmolarity and an unexplained metabolic acidosis. An elevation of 30–40 mOsm is equivalent to what level of ethylene glycol in the blood?

(d) Clinical laboratories use an assay of freezing point depression to measure serum osmolarity. What is the freezing point depression associated with the osmolarity of the serum samples in (a) through (c)?

8. In severe diabetes mellitus the blood sugar rises to very high levels often 700–1000 mg/100 ml. This is called a hyperosmolar state and is a medical emergency.

(a) What happens to the cells in the body during a hyperosmolar state?

(b) To preserve the cell volume of the neurons in the brain, these cells generate chemical species called “idio-osmols” that help resist the dehydrating effect of the hyperosmolar state. During treatment the blood sugar will fall to near normal levels and fluids are given to reverse the body’s dehydration. If the idio-osmols are still active, however, the brain can undergo life-threatening swelling. Explain why.

Part III
Building a Model of Biomolecular Structure

Chapter 14
Water: A Unique Solvent and Vital Component of Life

Contents

14.1 An Introduction to the Most Familiar of All Liquids

Water is a unique solvent by virtue of its extensive hydrogen-bonded character in bulk phase. Biological water has properties contributed by the properties of individual H_2O molecules and by the collection of H_2O molecules into a bulk phase with extensive intramolecular interactions. Therefore, on the basis of just our knowledge of statistical thermodynamics we would expect that the important properties of biological water will depend on the behavior of the collection of molecules or the entropic nature of water. However, biological systems are non-homogeneous, and non-bulk water plays a wide variety of physiological roles as well. Traditionally, the unusual solvent properties of water have been emphasized in studies of biophysical chemistry. An abundance of current investigation is expanding our understanding to the important role of water molecules and clustering in all aspects of biochemistry. These roles include hydrophobic interactions with small and large molecular structures as well as surfaces, surface chemistry interactions, structure–function relations in protein–protein interaction, mediation of ligand binding, biomolecular charge transfer, and modification of nucleic acid structure and interaction. This is

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_14,

certainly only a partial list and we will discuss aspects of many of these topics in the remainder of this book.

The percentage of water in biological components is substantial: 93% of the blood plasma is water and 80% of muscle and 60% of red cell mass is water. So ubiquitous is water in all of its phases that even the trained scientist takes much of its unique chemical quality for granted. “It is only water,” everyone of us has exclaimed. Water is unique in that it is the only naturally occurring inorganic liquid and chemical substance found in solid, liquid, and gas phases on the earth.

Water is continuously circulated between the earth’s hydrosphere, atmosphere, and terrosphere by the processes of evaporation and precipitation. This cycle is the *hydrologic cycle*. The magnitude of the water passing through of this cycle annually can be appreciated by examining Table 14.1. Less than 3% of the earth’s total water is fresh; 80% of the fresh water is frozen in polar and glacial ice. Ninety-five percent of the remaining fresh water is found in ground-water deposits. Still two-thirds of the planet is covered by water. Biological life not only lives in a watery environment (actually a wide variety of watery environments) but also participates in the hydrologic cycle. The water that passes through photosynthetic pathways and transpiration in plants measures nearly 10^5 km^3 annually. As we noted in Chapter 3, oxygen is the ultimate electron acceptor of the reducing equivalents drawn from reduced carbon by biosynthetic pathways. The fully reduced form of oxygen is H_2O and the fully oxidized form of carbon is CO_2 and these are the end products from the mitochondria in animal metabolism. Thus the average human produces 300 g of water daily de novo along with the liberation of 7600 kJ needed to produce ATP:

$$C_6\mathbf{H_{12}}O_6 + \mathbf{3O_2} + 3O_2 + 38\,\mathrm{ADP} + 38\,\mathrm{P_i} \rightarrow 6CO_2 + \mathbf{6H_2O} + 38\,\mathrm{ATP} \quad (14.1)$$

Table 14.1 The hydrologic cycle

Phase of cycle	Annual water volume (km^3)
Evaporation from the ocean surface	419, 000
Precipitation onto ocean	381, 000
Precipitation on land surfaces	107, 000
Run-off from land to ocean	38, 000
Evapotranspiration from land	69, 000

The biochemical mechanisms needed to perform this overall reaction, each use enzymes whose over structure is greatly affected by the aqueous environment in which they are formed and work. In addition, molecular water itself is a reactant in many of the steps of metabolic and catabolic pathways. Water is a participant in oxidation–reduction reactions as well as in hydrolysis and condensation reactions. Proper execution of these biochemical mechanisms depends on the properties of

H_2O, and the substitution of deuterium (H_2O) for H_2O leads to the derangement of the pathways and is toxic. Thus both the bulk and the molecular properties of water play a central role in biological life.

14.2 The Physical Properties of Water Are Consistent with a High Degree of Intermolecular Interaction

Water is an unusual compound whose physical properties set it apart from other liquids. Its properties include the following:

(1) Water has a high dielectric constant (Table 14.2). Water is therefore a very polar solvent, with the consequence that electrically charged molecules or particles are easily separated in its presence. Ionizations take place readily because the coulombic forces between opposite charges are weakened and more easily broken.

(2) Water has a very high heat capacity; a large amount of heat is needed to raise its temperature each degree K. In biological systems, this is an advantage, acting as a thermal buffer that prevents the temperature of a cell undergoing moderate metabolic activity from rising significantly. It has been speculated that if the heat capacity of water were not so high and warm ocean currents could not carry heat so efficiently, aquatic life would be profoundly affected and the oceans would be devoid of most fish. For example, the Gulf Stream is a stream of warm water 1/2 km deep and 200 km wide which carries an energy equivalent to 160 billion kg of coal burned per hour.

(3) Water has a high heat of vaporization (Table 14.3) as do other liquids capable of hydrogen bonding, such as ethanol and acetic acid. In contrast, non-hydrogen-bonded liquids such as hexane have lower heats of vaporization. This is the reason perspiration is such an effective method of cooling the body.

(4) Liquid water has a higher density than ice at standard temperature and pressure. There is a positive volume change upon freezing, which is why ice floats. If ice did not float, lake fish could not survive in lakes because the water would freeze from the bottom up. Furthermore, there is an insulating effect of the ice layer without which lakes would be too cold (i.e., completely frozen) to support life.

(5) Water has a high surface tension. This requires biological organisms to use detergent-like compounds to modify the surface tension. Lung surfactants are an example of these detergents. Lung surfactants significantly decrease the work needed to open the alveolar spaces and allow efficient respiration to take place. The absence of these surfactants leads to severe respiratory disease and often death.

(6) Water has a higher conductivity than might be expected on the basis of its ionization constant alone. The conductivity of ice is also high even though its

ionization constant is 1000-fold lower. This has been attributed to the presence of an ordered structure which allows the electric charge to be carried by a modified "brigade" mechanism.

Table 14.2 Dielectric constants of selected compounds

Substance	Dielectric constant (at 298 K)
Water	78.5
CH_3OH	32.6
CH_3CH_2OH	24.0
H_2S	9.3
C_6H_6	2.2
CCl_4	2.2
CH_4	1.7
Air	1.000059
Mica	5.4
Polystyrene	2.55

Table 14.3 Heats of vaporization for selected substances

Substance	Heat of vaporization (kJ/mol)
Water	40.7 at 373 K
Acetic acid	41.7 at 391 K
Ethanol	40.5 at 351 K
Hexane	31.9 at 341 K

14.3 Considering the Properties of Water as a Liquid

Taken together these properties make water unique in comparison to other liquids. But it is not so unique as to have nothing in common with other liquids. The properties of liquids can be analyzed from the viewpoint of intermolecular forces or structure. The structural view is emphasized when a liquid is considered as a perturbed solid. The liquid state results when the ordered crystalline structure of the solid state is disordered by thermal forces that increase the fluctuations of the molecules around their equilibrium positions. Thus, a liquid is a melted solid whose structure, though still present, is becoming more disordered. On the other hand, if a liquid is considered as a condensed gas, the role played by intermolecular forces is emphasized (Table 14.4). The liquid state results when gas molecules, which have no preferred position with respect to one another, are compressed to the degree that their motion is inhibited by the intermolecular forces that come into play as the molecules move into close proximity with one another.

Table 14.4 Liquids and intermolecular forces

Type of molecule	Dominant intermolecular force
Spherical monoatomic	van der Waals, short range
Homonuclear diatomic	van der Waals, short range, quadrapole
Metals, fused salts	Electrostatic, long range
Polar	Electric dipole
Associated	Hydrogen bonds, orientation dependent
Macromolecules	Intramolecular, hydrogen bonds, orientation

We can compare the melting and freezing points as well as the heats of vaporization of water to a series of isoelectronic molecules (all containing 10 electrons and 10 protons) as well as a series across the periodic table of equivalent hydrides. We find that H_2O exhibits anomalous behavior with surprisingly high points for melting, boiling, and heat of vaporization (Table 14.5). These findings indicate strong intermolecular forces. Thus water has significant intermolecular interactions. We will need to define the nature of these interactions. Our findings are striking when water is compared to neon and methane. These are very symmetrical molecules in terms of electronic charge distribution and hence only van der Waals interactions act between neon and methane. We anticipate water to be a molecule with electrical asymmetry and electrical dipole interactions. Does treating the intermolecular forces as dipole electrostatic interactions suffice? A polar molecule would be expected to have a high dielectric constant, but as Table 14.6 shows, other liquids made up of molecules with higher dipole moments do not have dielectric constants as high as water. We will explore these phenomena in some detail in Chapter 15 but, for now it is sufficient to point out that, if a group of dipoles are associated together in a liquid phase, the associated group has an effective dipole moment that is greater than the parts from

Table 14.5 The heats of vaporization as well as the melting and freezing points of water compared to its isoelectronic series and related hydrides

	Melting point (K)	Boiling point (K)	Heat of vaporization (kJ/mol)
Isoelectronic series			
CH_4	89	112	9.21
NH_3	195	240	23.22
H_2O	273	373	40.65
HF	181	292	30.20
Ne	24	27	1.74
Hydride series			
H_2O	273	373	40.65
H_2S	187	213	18.67
H_2Se	207	232	19.70
H_2Te	224	271	19.20

Table 14.6 Water has a higher dielectric constant than other liquids even though it does not have the highest dipole moment

Substance	Dipole moment	Dielectric constant (debye)
H_2O	1.85	80.1
C_2H_6OS (DMSO)	3.96	47.24
CH_3NO_2 (Nitromethane)	3.46	37.27
C_2H_4O (Acetaldehyde)	2.75	21.00
$(CH_3)2CO$ (Acetone)	2.88	21.01
H_2S	1.02	5.93

which it is made. Consequently we would expect water to be in some fashion a structured collection of dipoles. So far, we have been unable to pigeonhole water in terms used to classify liquids. The main explanation for water's singular properties is the presence of extensive hydrogen bonding between quadripolar water molecules. We will now examine the structure of monomolecular water, then explore its structure in bulk liquid phase.

14.4 The Structure of Monomolecular Water Can Be Described Using a Variety of Models

We can build up a model of the H_2O molecule in a fashion similar to our treatment of biological chemical bonding in Chapter 9. The Lewis structure for H_2O shows that a pair of covalent bonds links the hydrogen atoms to the oxygen with two lone pairs remaining on the oxygen. We next apply VSEPR theory and maximize the separation of all four sets of electron pairs. Since there are four sets of electron pairs, we predict the arrangement of electron pairs to be tetrahedral, but since two pairs are bonding and two pairs are non-bonding, the H_2O molecule will be classified as *angular*. Since non-bonding lone pairs occupy more space than bonding pairs, and bonding pairs will tend to move away from lone pairs and preferentially toward one another, the HOH angle would be expected to be less than the tetrahedral angle of 109.5°. X-ray diffraction studies show the HOH angle to be 104.27°.

We now apply valence bond theory. The electronic structure of the oxygen is $2s^2 2p_x^2 2p_y^1 2p_z^1$. Adding the electrons from the H1s orbitals leads to a sp^3 hybridization in which each of the unpaired electrons in the O2p orbitals pairs with an electron from the hydrogen. Two bonding orbitals and two orbitals containing lone pairs are formed. The asymmetry of these orbitals also leads to the conclusion that the observed angle should be less than 109.5° (Fig. 14.1). The sp^3 hybridization model is a powerful cartoon and as Fig. 14.4 shows, the description of the lone pairs as sticking out from the oxygen like a pair of ears from a rabbit should be resisted. We will see why in the next paragraphs.

Penultimately, we consider the LCAO-MO model which will allow us to account for the polar covalent bonds of the H_2O molecule. The LCAO-MO model takes into

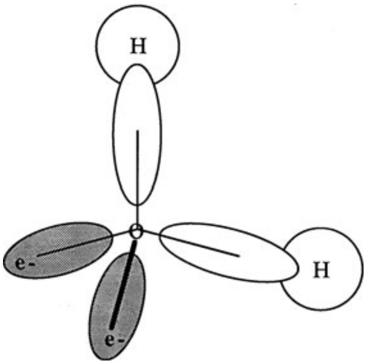


Fig. 14.1 The orbitals are arranged so that the protons lie in one plane of the terahedron and the lone pairs in a plane perpendicular to the proton plane

account the electronegativities of the elements to calculate the percentage of time the electronic charge spends in the vicinity of one or the other of the atoms in a heteronuclear molecule. The Pauling electronegativity of H is 2.1 and that of O is 3.5. We would expect the atomic orbitals with the lowest energies to be located near the oxygen atom and the antibonding orbitals to be found near the hydrogen atoms. The statistical consequence will be a greater time spent by the electronic charge near the oxygen atom resulting in a partial negative charge near the oxygen and a corresponding partial positive charge near the hydrogens. We build the molecular orbital picture from a linear combination of all of the atomic orbitals without attempting to construct bonds between atoms (Fig. 14.2). H_2O is a triatomic heteronuclear hydride, and its molecular structure can be considered along with BeH_2,

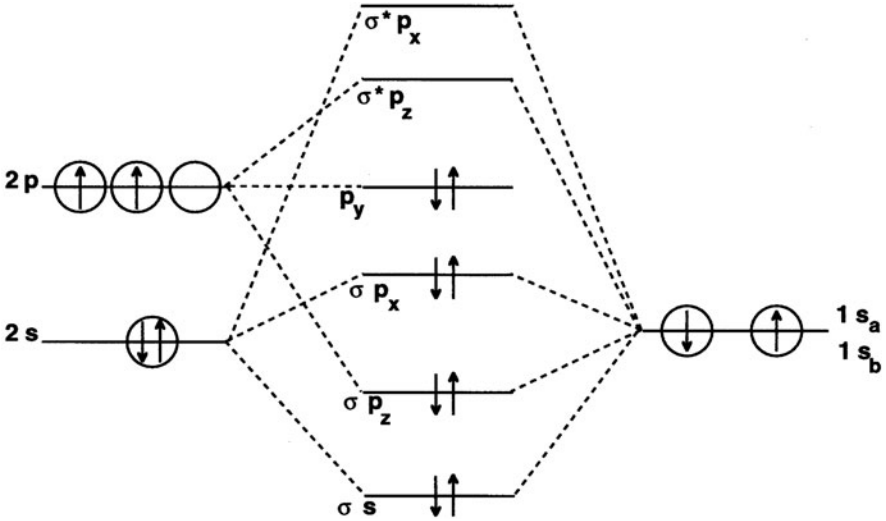


Fig. 14.2 Molecular orbital diagram of water

BH_2, CH_2, and NH_2. This series of molecules forms molecular orbitals derived from two H1s, one O1s, and three O2p orbitals. The O1s orbital is held closely to the oxygen nucleus and is a non-bonding orbital leaving the remaining four atomic orbitals to form four molecular orbitals. As Fig. 14.3 shows the orbital energies of the XH$_2$ series vary as a function of the HXH bond angle which accounts for the bent structure of H_2O.

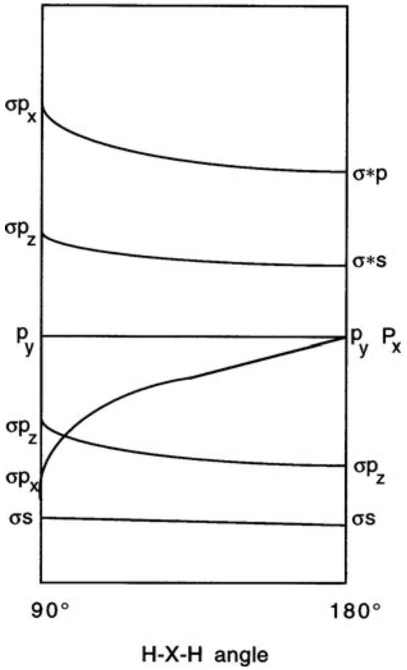


Fig. 14.3 Adding electrons sequentially to the molecular orbital energy-level diagrams of the prototypic bent XH$_2$ molecule demonstrates why the structure of H_2O is bent. Consideration of this diagram would predict that BeH_2 is linear; BH_2 is bent in the ground state and linear in the excited state; NH_2 is bent. CH_2 is linear because it is stabilized by having unpaired electrons in non-bonding 2*px* and 2*py* orbitals. These predictions are all borne out by experiment

The most complete description of a single water molecule would be given by solution of a quantum mechanical treatment of the H_2O molecule, for example, the Hamiltonian for the water molecule:

$$\mathcal{H} = E_n + E_e + U(\mathbf{r}, \mathbf{R}) \tag{14.2}$$

Here the first and second terms account for the kinetic energy of the three nuclei and the 10 electrons in the molecule, respectively. The last term is the potential energy function for the electrostatic interaction between all pairs of particles whose coordinates are **r**, **R**. Generally the Born–Oppenheimer approximation is applied and the first term is dropped. The electronic motions in a fixed nuclear force field are then calculated. Solution of the nuclear Schrödinger equation predicts theoretical, vibrational and rotational motions, and measurements by infrared spectroscopy confirm the vibrational modes that are calculated. The equilibrium geometry of the molecule has an O–H bond length of 0.0958 nm and the HOH bond angle is 104.27°.

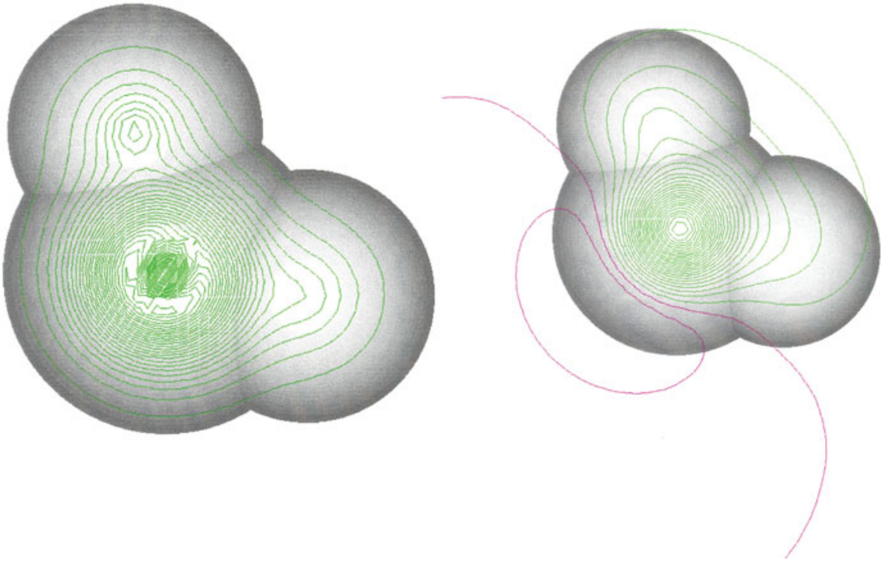


Fig. 14.4 The calculated charge distribution for a water molecule is shown in the *upper* frame with the resulting electrostatic field in the lower. The positive field lines are pointing to the lower left corner of the page while the negative field is pointing to the upper right-hand corner (Calculations are by an intermediate neglect of differential overlap semi-empirical method)

Figure 14.4 shows a calculation of the electronic charge distribution and the resulting quadripolar electrostatic field though it can be appreciated that the electronic distribution shows a broad maximum on the "back" of the oxygen atom. This has implications for the number and "occupancy" of hydrogen bonds formed in liquid water, a topic that we will discuss shortly.

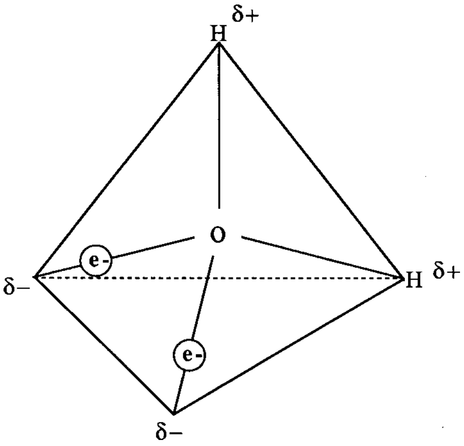


Fig. 14.5 Tetrahedral bonding arrangement proposed by Bjerrum, where δ+ is a partial positive charge and δ− is a partial negative charge

The four point charge model described by Bjerrum remains a useful model of the water molecule though today its details have been modified by quantum mechanical treatments. This model (Fig. 14.5) places the oxygen atom at the center of a tetrahedron, and fractions of charge are placed 0.1 nm from the center. The molecule is given the same van der Waals radius as its isoelectronic cousin neon, 0.282 nm. All of these models show a separation of charge in the water molecule thus predicting a molecule with a significant permanent dipole moment. This dipole nature will be important as will be evident in the next several chapters. (Though the theoretical models and experiment show that the water molecule is an electrical quadripole, we will restrict ourselves to the more convenient abstraction of a dipole in this book).

14.5 The Capacity of Water to Form Hydrogen Bonds Underlies Its Unusual Properties

Given the structure and polar nature of the covalent bonds in the water molecule, it is easy to imagine an interaction between the non-bonding electron-rich orbitals and the relatively electron-deficient hydrogen nuclear centers (Fig. 14.6). Water dimers can be found to associate through hydrogen bonding, and theoretical calculations predict a linear hydrogen bond with molar dissociation of 20–35 kJ and equilibrium O···O distance of 0.26–0.30 nm. If the hydrogen bond is bent less than 25° there does not seem to be a significant change in the dimer bond energy. An important feature of dimer formation is the charge displacement that occurs which can be viewed as a covalent contribution to the interactional energies (Fig. 14.7). This charge displacement acts to stabilize larger aggregates of water molecules and leads to the favorable energetics, i.e., more probable association, of a single molecule with an aggregate group than with another single molecule to form a dimer. Therefore, the aggregation of water molecules into larger aggregates is *cooperative*.

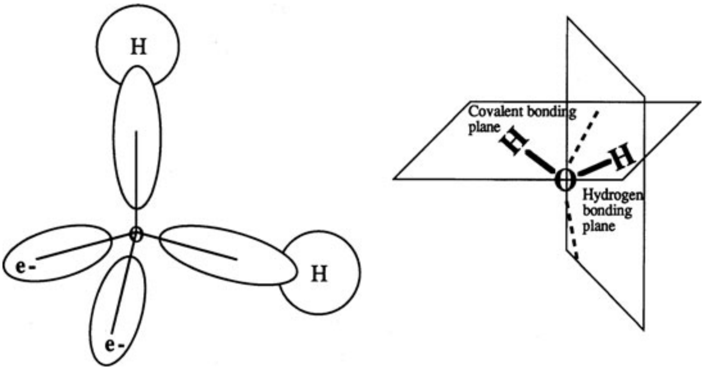


Fig. 14.6 Geometry of the molecular orbitals of water; the axis is perpendicular to the page

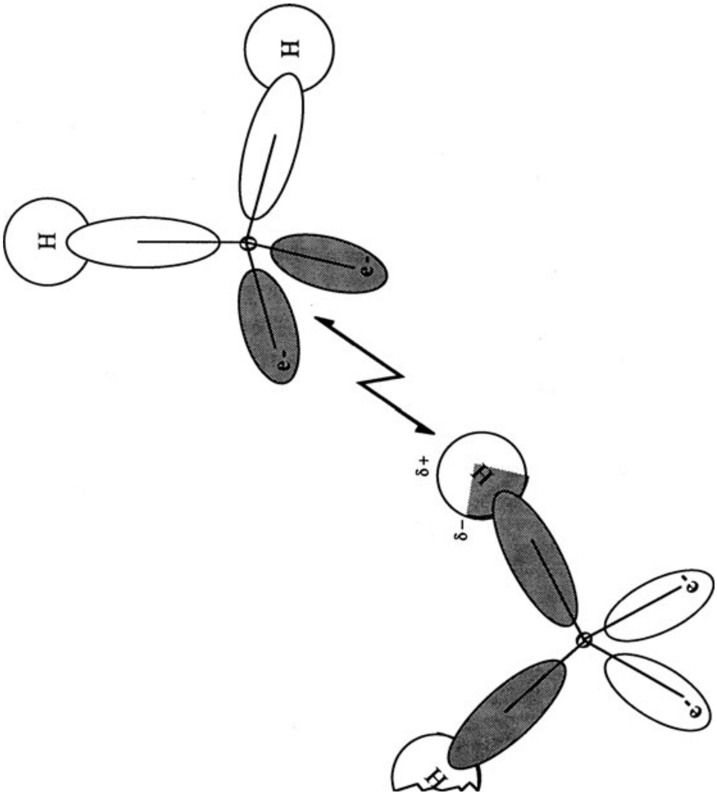


Fig. 14.7 The charge distribution that leads to formation of a hydrogen bond, showing the partial positive charge on the hydrogen atom due to the strong electronegativity of the covalently bonded atom. This partial positive charge interacts with the electron density of the unbonded pair of electrons on the second electronegative species

Hydrogen bonds arise because the structure of the water molecule favors such interactions, as its molecular orbital representation indicates. In the case of water, hydrogen bonding is strongly favored because each proton that is covalently bonded to the strongly electronegative oxygen atom finds an unbonded electron with which to interact in a one-to-one relationship. Because of this one-to-one proton-to-electron relationship, the orbital structure dictates that each oxygen atom on the average is involved in four bonds to hydrogen atoms. Two bonds are covalent and two are non-covalent hydrogen bonds. X-ray diffraction of liquid water shows a major peak at 0.28 nm which corresponds to the distance between oxygen atoms in the tetrahedral hydrogen-bonded water structures, confirming the existence of the predicted hydrogen-bonded structure (Fig. 14.8).

Although the unusual properties of H_2O were well known, experimental evidence of the extensive existence of hydrogen bonds came when spectral techniques

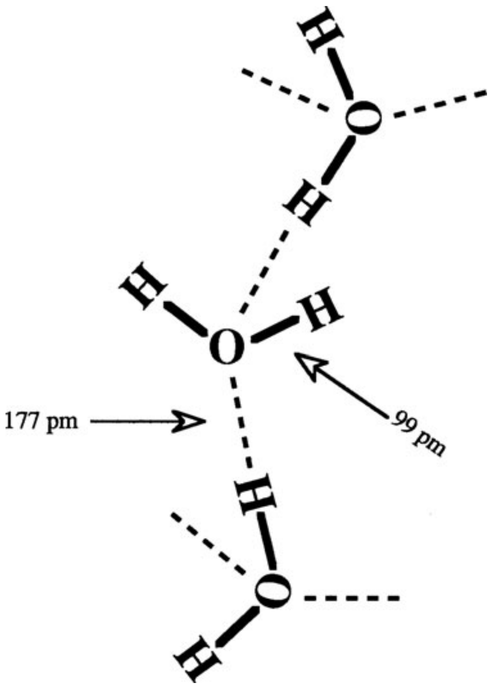


Fig. 14.8 Geometry of the hydrogen bonding. The water molecules are arranged so that each oxygen has a pair of covalent bonds and a pair of hydrogen bonds

were developed which permitted comparison of normal –OH bonds (e.g., in alcohols) with –OH bonds in water. These studies were largely based on the effect of hydrogen bonds on the vibrational transitions shown in Fig. 14.9. The energies associated with such transitions fall in the infrared spectral region. Molecules which do not interact with each other, such as molecules of water vapor at very low pressure, undergo the normal vibrational modes shown in Fig. 14.9(a–c). The corresponding infrared spectra of liquid water shows each of these vibrational modes, especially deformation, to be perturbed, thus providing evidence that the water molecules participate in extensive mutual hydrogen bonding (Fig. 14.9(d–f)). These types of studies originally were made difficult by the very broad infrared bands associated with the H_2O molecule but became possible with the availability of D_2O, which permitted determination of the shifts in absorbance peaks caused by hydrogen bonding. Such studies led Pauling to postulate that the hydrogen bond was a major interaction which played a critical role in water structure and also in the structure and function of biological macromolecules.

The importance of hydrogen bonding in water can be emphasized by comparing the properties of H_2O and H_2S. Hydrogen sulfide is a compound which might, by virtue of the location of S below O in the periodic table and its structure being almost identical to that of H_2O, be expected to have similar properties to water (Fig. 14.10). The O–H bond in H_2O is stronger than the S–H bond in H_2S, since the respective lengths of these bonds are 0.099 nm and 0.134 nm. There is also no

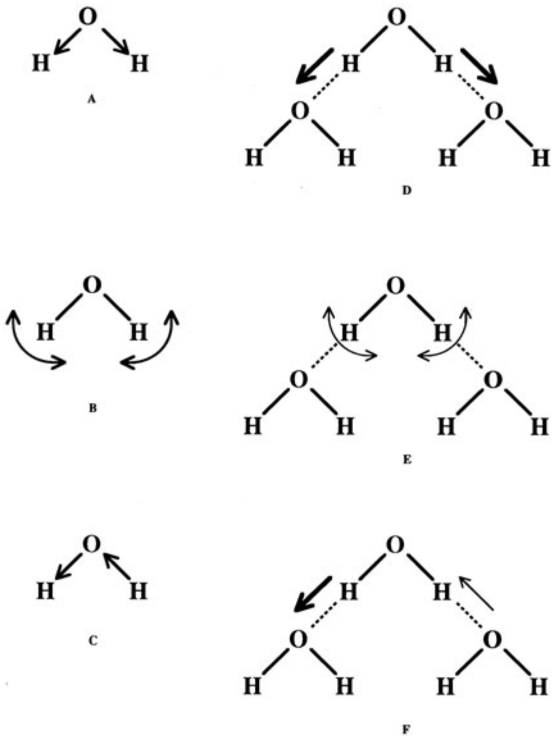


Fig. 14.9 Vibrational modes of O–H bonds are affected by the presence of hydrogen bonds. **(a–c)** are the vibrational motions without hydrogen bonding: (**a**) the symmetric stretching; (**b**) the deformation of the bonds; (**c**) the asymmetric stretching. In **(d–f)** hydrogen bonds affect each of these vibrational modes: (**d**) the symmetric stretching is enhanced and needs less energy; (**e**) the deformation of the bonds is inhibited and needs more energy to occur; (**f**) the asymmetric stretching is partly enhanced and partly inhibited

hybridization of atomic orbitals in H_2S. The “tetrahedral” angle between bonds in H_2O is 104°30′, whereas the angle in H_2S is 92°20′, an angle which cannot fit into a tetrahedral array and which prevents the formation of any further bonds to the S atom. Both molecules are polar, but the dipole moment for H_2O is considerably larger (1.8 debye for H_2O, 1.1 debye for H_2S). As Table 14.7 indicates, there are significant differences between the various thermodynamic parameters of these two compounds which in the absence of hydrogen bonds might have been expected to be rather similar. The main general observation that can be made about the properties

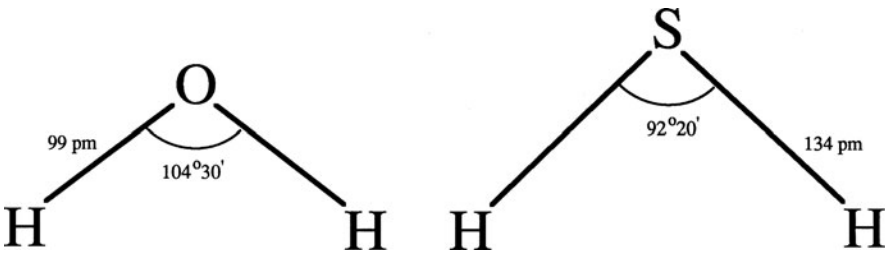


Fig. 14.10 Structures, bond lengths, and angles for H_2S and H_2O

Table 14.7 Comparison of thermodynamic parameters of water and hydrogen sulfide

Property	H_2O	H_2S
$T_{\text{solid–liquid, 1 atm}}$	273 K	187 K
$T_{\text{liquid–gas, 1 atm}}$	100 K	213 K
Density$_{\text{solid}}$ at T_{m}	0.9998 kg/dm^3	1.80 kg/dm^3
$C_{p,\text{ liquid}}$	76.02 J/mol/K	36.1 J/mol/K
$C_{p,\text{ gas}}$	36.44 J/mol/K	41.56 J/mol/K
$\Delta H_{\text{solid–liquid, 1 atm}}$	6.003 kJ/mol	2.386 kJ/mol
$\Delta H_{\text{liquid–gas, 1 atm}}$	40.656 kJ/mol	18.747 kJ/mol
Conductivity at $T_{\text{liquid–gas, 1 atm}}$	$4\times10^{-10}\ \Omega^{-1}\ \text{m}^{-1}$	$1\times10^{-13}\ \Omega^{-1}\ \text{m}^{-1}$

and the thermodynamic variables given in Table 14.7 is that more energy is required to heat, melt, or vaporize 1 mol of H_2O than 1 mol of H_2S. This is attributed to the strong tendency of H_2O to form hydrogen bonds with other water molecules, while H_2S does not form hydrogen bonds.

14.6 The Structure and Dynamics of Liquid Water Results in "Ordered Diversity" That Is Probably Distinct from Ice

Our emphasis on the structure of the molecule and its intermolecular forces is consistent with a view of liquid water as a condensed gas. However, the hydrogen bonding gives rise to a highly stabilized solid crystalline structure, and an appreciation of the structure of ice is valuable since liquid water has many structural characteristics of crystalline ice. The degree of the preserved structure is due in great degree to the hydrogen bonding between the water molecules. There is considerable agreement upon the most probable structure for *crystalline water* (ice) at 1 atm pressure and 273 K, shown in Fig. 14.11. We can treat the crystal structure as if

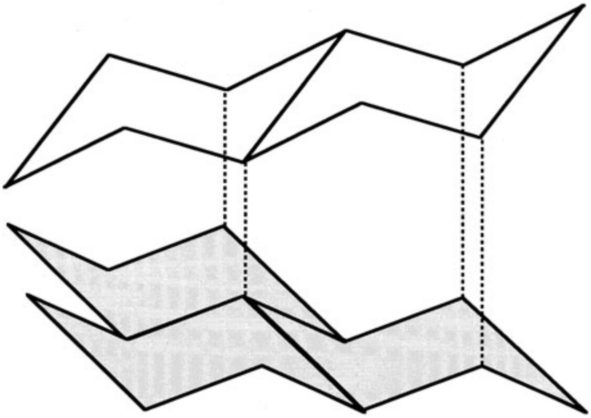
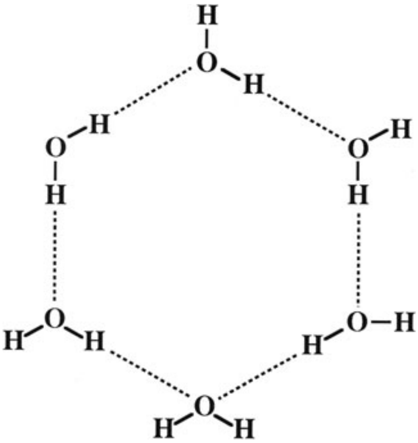


Fig. 14.11 Structure of ice at 1 atm and 273 K

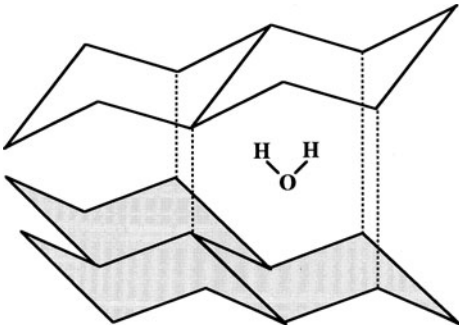
four bonds point toward the corners of a tetrahedron resulting in tetrahedral arrays of hydrogen-bonded water. The cross section of these bonded arrays creates a hexagon as seen in Fig. 14.12. The structures are reminiscent of the chair configuration found in cyclic six-carbon hydrocarbon rings.

Fig. 14.12 Cross section of the tetrahedral array shows the hexagonal structure of ice



The hexagonal arrays of H_2O are quite large with ample amounts of empty space; the hexagons in ice as depicted in Fig. 14.11 are empty. These spaces within the hexagons could be filled by small molecules such as monomolecular water (Fig. 14.13) or other small uncharged molecules. Under high pressure, at low temperatures, an entire second hexagonal array can be created within the spaces of the first. Any array having molecules interspaced within the ordered structure will have a higher density than the comparable structure with empty spaces. In other words, H_2O molecules can penetrate into the hexagonal structure of ice under conditions of greater pressure or into the less regular, but still hexagonal, structures of liquid H_2O.

Fig. 14.13 A small molecule can fit inside the spaces of the hexagonal arrays of water



A rigorous discussion of the various theories of liquid water structure is beyond the intended scope of this text but references can be found at the end of the chapter. There is still controversy as to the actual structure of liquid water. A number of models of the structure of liquid water have been proposed, differing largely in the mathematical treatments. Most of them treat liquid water as a mixture of distinct states of organization or as a continuum of structures. All models share the feature of being based on a high degree of hydrogen bonding between water molecules and take into account the strong dependence of the hydrogen-bond energy on the configuration of the bond. If there is bending of the hydrogen bond so that the atoms forming the bond are not collinear, then there is a loss of bond energy. Consideration of the effects of the interaction of the water dipoles with one another is also a general feature of the theories. Experimental evidence is growing that may bring some resolution to these modeling controversies. There is now growing evidence from synchrotron studies that while in ice each water molecule is surrounding by four other molecules in a tetrahedral arrangement in the liquid state there are only two other connections. Thus a diversity of other structures such as strongly hydrogen-bonded rings or chains is embedded within a disordered cluster network. This cluster network is likely connected mainly by weak hydrogen bonds.

This hydrogen-bonded structure leads to a bulk water state with a very high entropy and to a much lower entropic (and therefore less favorable) state when single molecules are isolated from the bulk. The high entropy of water is a result of the equivalence of the hydrogen bonds throughout the bulk phase. Because of this equivalency, there is an enormous degeneracy in the possible energy states that a particular water molecule may occupy throughout the bulk phase. As a result of this degeneracy, a large number of microstates, and hence a large entropy, can be postulated. In entropic terms then, the hydrogen bonding in water does not restrict choice, but rather maximizes choice for the water molecules. In other words, a particular water molecule in the bulk structure cannot be identified with certainty. It is important to recognize the unique aspect of this entropic effect, because when water molecules are removed from the bulk, or restricted in their ability to associate with the bulk, the possible number of microstates is sharply decreased, and the entropy of the system falls. This effect will be important in biological solvent–solute interactions and will be discussed further in coming chapters.

The salient features of these models can be explored by considering the behavior of proton conduction in ice through a brigade-type mechanism as proposed by Klotz (cf. Chapter 23). Since all water molecules are identical, the hydrogen bonds in an array can be exchanged to include a new water molecule with virtually no loss in energy, as long as the net number of bonds and their environment remain constant. In the Klotz model, a specific proton does not need to move physically in order to propagate a current. Instead, hydrogen bonds break and then reform successively like a group of dominoes falling, with the net result being the apparent net movement of a proton. This unique capability accounts for the high electrical conductivity of ice even though the degree of ionization is 1000-fold lower than that of liquid water ($\mathrm{p}K_{\mathrm{ice}} = 10$, $\mathrm{p}K_{\mathrm{liquid\ water}} = 7$). The structure of ice and water facilitates the charge transfer, and the conductivities of ice and of liquid water are hence almost equal. The case of proton conduction will be considered in more detail in Section 14.6.

The Klotz model proposes a single class of H_2O molecules. This means that all the molecules are considered to behave identically and to have identical thermodynamic properties. A continuum of a hydrogen-bonded species exists in which bonds are constantly being made and broken between indistinguishable molecules. This type of structural model satisfactorily accounts for the observation by nuclear magnetic resonance spectroscopy that there are resonance peaks only for a single type of hydrogen-bonded species. Mixtures of defined clusters or aggregates with non-hydrogen-bonded species (i.e., two different types of H_2O, bonded and non-bonded) are not compatible with the data. The flickering cluster of Klotz's model is characterized by the constantly changing association and reassociation of groups of 50 to 70 water molecules. The time of association is very short in the liquid state, approaching 10^{-12} s, and shortens as the temperature increases.

The debate over the various models of liquid water structure is ongoing. For the purposes of this text, it is sufficient to regard all models as essentially "mixture" models, in which the structure of water is treated as a dynamic, fluctuating assembly of water molecules at different states of hydrogen bonding. Bulk water almost certainly contains a diversity of structures in the liquid phase. The differences between the models lie in the size, extent, and degree of deformation of the hydrogen-bonded component(s) of the mixture, since, as Franks has shown, there is cooperativity between hydrogen bonds. Overall there is consensus that in the liquid state water is in a dynamic state of associated clusters.

In summary then, there is greater Brownian motion in liquid water, and consequently, there is less organization and greater translational rearrangement than in ice. Hydrogen bonds in liquid water are weaker than those in ice because the interaction between adjacent hydrogen bonds is statistically more likely to be absent, rendering each individual hydrogen bond weaker. The nature of pure water and its properties contribute to the formation of hydrogen bonds. A single kind of hydrogen bond between identical H_2O molecules is implicated in the properties of crystalline and to some degree liquid water. As will be seen in the following chapters, the role that this structured water plays in solvent–solute and solute–solute interactions is significant and important to understand.

14.7 Hydrophobic Forces Reference Interactions Between Water and Other Molecules

The special role played by water as a solvent and a solute in biological systems has lead to the perspective that water itself is a reference against which other substances should be measured. Thus substances are classified as hydrophobic and hydrophilic recognizing the phenomenology that when mixed with water some materials will associate with the water phase and some will not. Since the likelihood of the state of association can be expressed in terms of a free energy of hydrophilicity, the hydrophobic/hydrophilic phenomenology has been further extended to include "hydrophobic forces." But hydrophobic forces are in fact a *pseudoforce* just like the centrifugal force (see Appendix N) and the diffusion forces (Chapter 22) with whose effects we are familiar. This is a very important epistemological distinction between "real" and "pseudo" forces. Though all forces require an interaction between the observer and the observed to be known, forces like electrical and mechanical ones result from specific interactions between individual entities and are measurable between these individuals. A real force is a characteristic of the entity within a system under study. Pseudoforces are also measurable in terms of interactional events but they are not properties of individual entities but rather are properties of the system. Because a pseudoforce depends on the system configuration, its behavior is sometimes confusing and often paradoxical. Hydrophobic forces often manage to confuse and bewilder the student. However with these introductory comments, these water-related forces should be more clearly understood.

Hydrophobicity is best regarded as reflecting two separate phenomena. The first derives from the strong hydrogen bonding nature of water that makes it a highly associated liquid. In this chapter we have introduced the structure of water that results from the hydrogen bonding. We will see in particular in detail in the following chapters that when the water phase and a non-polar solute are forced together, the water molecules make all attempts possible to preserve their maximal hydrogen bonding interactions. The water will try to find an orientation in which hydrogen bonds can be preserved by reorienting themselves around the solute. So while the enthalpy of the interaction, which primarily due to the hydrogen bonds, is preserved (in fact the number of hydrogen bonds is often slightly increased from an average of 3–3.5 to about 4) the effect is that the water itself is more structured around the solute and therefore distinctly different compared to other water molecules in the bulk. This more highly ordered structure is entropically unfavorable and thus the system is energetically inclined to avoid moving in such a direction. This phenomenon is called *hydrophobic hydration* or *solvation* and the overall effect of non-polar substances being so sparingly soluble in water is due to this *hydrophobic effect*. The dominance of the entropic factor can be appreciated by considering the $\Delta G_{\text{transfer}}$ of *n*-butane to water which is equal to +24.5 kJ/mol. For this transfer ΔH is –4.3 kJ/mol, while $-T\Delta S = 28.7$ kJ/mol. An important point can be made about these numbers. There is an attractive force (represented by the negative enthalpy) between water and hydrocarbon molecules. Most of this is due to the attractive dispersion forces and some to the increased hydrogen bond number (the H-bonds are not of different energy). Thus the overall $+\Delta G$ occurs not because of repulsion between the water and hydrocarbon or non-polar molecule but rather due to the overwhelming tendency for water molecules to prefer their own company to the exclusion of even positive interactions with other entities.

The free energy of transfer for hydrocarbons is roughly proportional to the surface area of the molecules and can be calculated to be in the range of 40–50 mJ/m^2. This value is consistent with the measured surface tension in hydrocarbon–water systems which is largely an entropic effect. We know that the high surface tension of the water–air interface (72 mJ/m^2) is largely an entropic effect due to the hydrogen-bonded nature of water avoiding interaction with a non-hydrogen bonding medium (the air). As unfavorable as these events that force water into highly structured interfacial arrangements are, at least some hydrogen bonding and choice is left to the water molecules. If a water molecule is isolated into a region where no water network at all is available, the energetic cost is even higher. We will see how such factors will drive the self-assembly of lipid membranes and folding events in proteins.

The description of non-polar molecules that tend to associate with one another as having a hydrophobic character is curious because these molecules will associate in completely water-free environments. This association is almost wholly due to the van der Waals or dispersion forces that we have studied. However there are phenomena involving these non-polar/non-polar interactions that need consideration. The *hydrophobic interaction* is related yet distinct from the hydrophobic effect. This term refers to the phenomena in which the association of non-polar molecules is stronger in an aqueous environment than in an aqueous free system. For example, the surface tension of saturated hydrocarbons in air range from 15 to 30 mJ/m^2 yet the hydrocarbon surface tension at the interface with water is on the order of 40–50 mJ/m^2 as we noted above. In addition the van der Waals energy between two methane molecules in free space is -2.5×10^{-21} J but if those same molecules interact in water the interactional energy increases over fivefold. It is these increased energies that lead to the thought that a hydrophobic bond existed. However these interactions are due to the hydrocarbon–water system and have at their foundation the same entropic etiologies as the hydrophobic effect.

Further Reading

Ball P. (2008) Water as an active constituent in cell biology, *Chem. Rev*., **108**:74–108.

Braum C.L. and Smirnov S.N. (1993) Why is water blue? *J. Chem. Educ*., **70**:612–614.

Eisenberg, D. and Kautzmann W. (1969) *The Structure and Properties of Water.* Oxford University Press, Oxford.

Franks F. (ed.) (1973) *Water, A Comprehensive Treatise,* Volumes 1–6. Plenum Press, New York.

Franks F. (1983) *Water.* The Royal Society of Chemistry, London.

Franks F. (1990) *Water Science Reviews.* Cambridge University Press, Cambridge.

Franks F. and Mathias S.F. (1991) *The Biophysics of Water*: *Proceedings of a Working Conference Held at Girtin College*. Cambridge Books on Demand, Ann Arbor, MI.

Tanford C. (1980) *The Hydrophobic Effect.* Wiley, New York.

Articles on Water Structure

Bryant R.G. (1996) The dynamics of water-protein interactions, *Annu. Rev. Biophys. Biomol. Struct*., **25**:29–53.

Colson S.D. and Dunning T.H. (1994) The structure of nature’s solvent: Water. *Science,* **265**:43–44.

Cruzan J.D., Braly L.B., Liu K., Brown M.G., Loeser J.G., and Saykally R.J. (1996) Quantifying hydrogen bond cooperativity in water: VRT spectroscopy of the water tetramer. *Science*, **271**:59–62.

Elolaa M.D. and Ladanyib B.M. (2006) Computational study of structural and dynamical properties of formamide-water mixtures. *J. Chem. Phys.*, 125:184506, doi:10.1063/1.2364896.

Forslind E. (1971) Structure of water. *Q. Rev. Biophys*., **4**:325–363.

Isaacs E.D., Shukla A., Platzman P.M., Hamann D.R., Barbiellini B., and Tulk C.A. (1999) Covalency of the hydrogen bond in ice: A direct X-ray measurement. *Phys. Rev. Lett.*, **82**:600–603.

Jayaram B. and Jain T. (2004) The role of water in protein-DNA recognition. *Annu. Rev. Biophys. Biomol. Struct.*, **33**:343–361.

Kusalik P.G. and Svishchev I.M. (1994) The spatial structure in liquid water. *Science*, **265**: 1219–1221.

Levy Y. and Onuchic J.N. (2006) Water mediation in protein folding and molecular recognition. *Annu. Rev. Biophys. Biomol. Struct*. **35**:389–415.

Liu K., Brown M.G., Cruzan J.D., and Saykally R.J. (1994) Vibration-rotation tunneling spectra of the water pentamer: Structure and dynamics. *Science*, **271**: 62–64.

Pribble R.N. and Zwier T.S. (1994) Specific infrared spectra of benzene-$(H_2O)_n$ structures. *Science*, **265**:75–79.

Savage H. and Wlodawer A. (1986) Determination of water structure around biomolecules using X-ray and neutron diffraction methods. *Methods Enzymol*., **127**:162–183.

Strauss H.L. (1986) Vibrational methods for water structure in nonpolar media. *Methods Enzymol*., **127**:106–113.

Svishchev I.M. and Kusalik P.G. (1996) Electrofreezing of liquid water: A microscopic perspective. *J. Am. Chem. Soc*., **118**:649–654.

Tanford C. (1996) How proteins chemists learned about the hydrophobic factor. *Protein Sci*., **6**:1358–1366.

Wernet Ph., Nordlund D., Bergmann U., Cavalleri M., Odelius M., Ogasawara H., Näslund L.Å., Hirsch T.K., Ojamäe L., Glatzel P., Pettersson L.G.M., and Nilsson A. (2004) The structure of the first coordination shell in liquid water. *Science*, **304**:995–999.

Problem Sets

1. From your knowledge of water structure, explain why ice floats. Would you expect ice and water to have the same structure in Boston as at the bottom of the ocean? How would these structures differ? What implications does this have for the biochemicals that might make up a deep-water fish?
2. In 1997, great excitement surrounded the presumed existence of water on one of the moons of Jupiter. This moon was thought to have enough heat generated from its volcanic activity that water would be melted underneath a mantle of ice. What is the likely heat capacity, dielectric constant, and heats of formation and vaporization for this Jovian water when compared to terrestrial water?
3. A Jovian enzyme is discovered by a space probe that appears to act by adding a water molecule to an ester bond (hydrolysis). Assuming that the enzyme maintains its structure when brought back to Earth, comment on the relative activity of the hydrolyzing water in the terrestrial environment.

Chapter 15
Ion–Solvent Interactions

Contents

15.1 The Nature of Ion–Solvent Interactions Can Be Discovered Through the Progression of Inquiry

We will now apply our knowledge of modeling to a system of biological importance. This model-building exercise is a reasonably detailed examination of the forces that exist between ions and a solvent. The treatment of these forces will lead to an understanding, in molecular terms, of the deviations from the ideal, characterized empirically by the activity coefficient. Initially, a continuum model that neglects any knowledge of the molecular structure of water will be proposed. The empirically inspired corrections to the model will point the way toward a more detailed examination of ion–solvent interactions. Initially, the behavior of ion–solvent interactions will be described in terms of simple electrostatics, but this will turn out to

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_15, © Springer Science+Business Media, LLC 2010

be a crude approximation. As further detail about the solvent and its interaction with ions is sought, however, it will become apparent that the structure of water has important effects in modifying the electrostatic environment. In considering the dielectric constant and the Kirkwood equations, the effects of water structure on the electrostatic forces are described in some detail, an exercise that is important when considering the more complex relationships in cellular systems.

In studying this section, we will endeavor to gain a reasonable knowledge about the relationship of an aqueous solvent to ions and dipolar molecules and also to begin to achieve mastery of modeling techniques as a method of scientific research.

15.2 The Born Model Is a Thermodynamic Cycle That Treats the Interaction Energy Between a Simplified Ion and a Structureless Solvent

A set of assumptions will now be set forth that may seem too simplistic initially, but there is value in understanding the model that can be built with these assumptions. Simple models are usually most appreciated when the formulation becomes more accurate and also more mathematically complex. The system for first consideration will be the one proposed in the early part of this century by Max Born and is called the *Born model*. The model is based on several assumptions (Fig. 15.1):

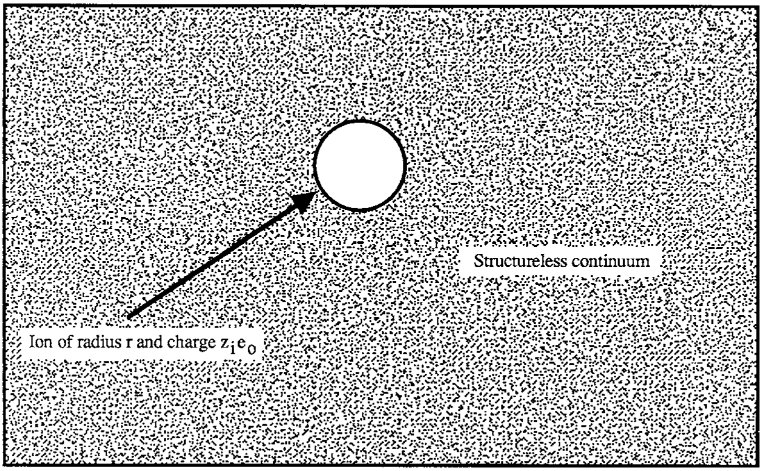


Fig. 15.1 The basic assumptions on which the Born model is based: (1) that the ion under consideration can be represented by a sphere of radius r_i and charge $z_i e_o$ and (2) that the solvent into which it is dissolved is a structureless continuum

(1) The ion may be represented as a rigid sphere of radius r_i and charge $z_i e_o$, where z_i is the charge number and e_o is the charge on an electron.
(2) The solvent into which the ion is to be dissolved is treated as a structureless continuum.
(3) All interactions are electrostatic in nature (i.e., $F \propto \frac{1}{r^2}$).

When the model's predictive power is assessed by experiment these are the assumptions that will be tested. A priori then, this model should permit an understanding of the relationship between an ion and its solvent, in our case water.

15.2.1 Building the Model

The ultimate picture of this model is the formation of a solution of ions, represented as rigid spheres of a particular charge, by floating them in a continuum. Where do these charged spheres come from? The source of the ions is really a matter of convenience, and it is most convenient to consider the problem as the movement of a charged sphere (the ion) from a vacuum into the continuum. The attention paid earlier to the concept of a thermodynamic cycle will be valuable here. If the work done each step of the way in the movement of the charged sphere from vacuum to continuum and back to vacuum is summed, a cyclic integral will result. Since the entire nature of the ionic character is to be given in this model by the fixed charge on a rigid sphere, the work functions and ion–solvent interactions will be defined strictly by electrostatic forces. The goal is to write a thermodynamic cycle that is convenient to solve, so the following cycle is proposed (Fig. 15.2):

(1). Start with a charged sphere in a vacuum.
(2). Discharge the sphere so that it is electrically neutral; this will be the *work of discharging*.
(3). The discharged sphere can now be freely slipped into the continuum. Since the only interactions allowed are due to electrostatic forces, without charge on the sphere, no interactional force exists. Although equal to zero, this is the *work of mass transfer*.
(4). Once inside the continuum, the sphere is charged back up to its original charge, $z_i e_o$. This will be the *work of charging*.
(5). The goal is to design a thermodynamic cycle so the charged sphere is now removed from the continuum and brought back to its starting place in the vacuum. The work necessary to do this is the *free energy of the ion–solvent interaction*, $\Delta G_{\text{ion}-\text{solvent}}$ Since the work to remove the sphere from the continuum into a vacuum is equal to, but opposite in sign to, the work necessary to move the sphere into the continuum, the free energy is $-\Delta G_{\text{ion}-\text{solvent}}$ $(-\Delta G_{\text{i}-\text{s}})$.

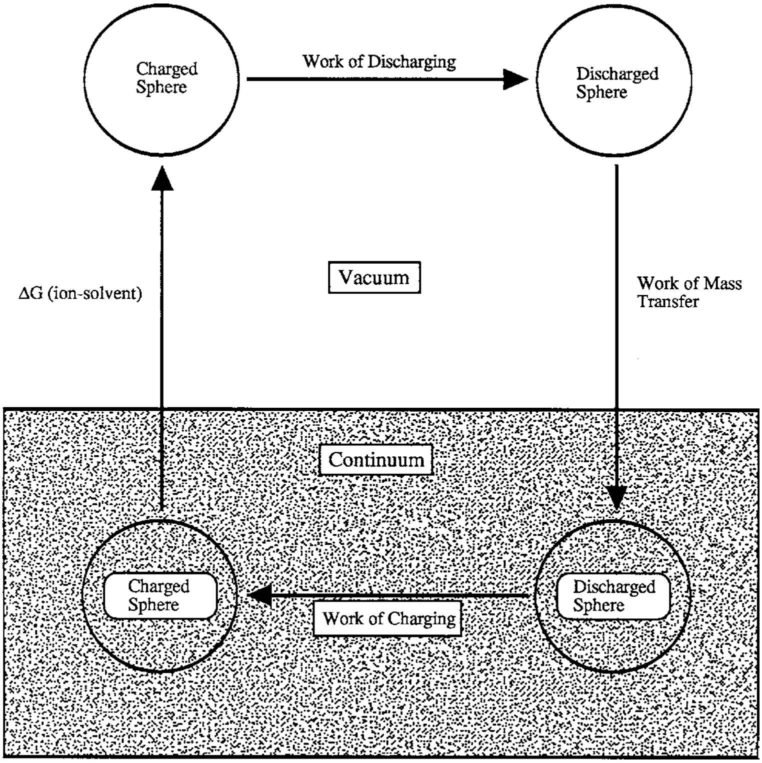


Fig. 15.2 Schematic of the Born charging cycle

This cycle can be written out as follows:

$$w_{\text{discharge}} + w_{\text{transfer}} + w_{\text{charging}} + (-\Delta G_{i-\text{s}}) = 0 \qquad (15.1)$$

Since $w_{\text{transfer}} = 0$ (cf. step 3 above), this can be algebraically manipulated:

$$w_{\text{discharge}} + w_{\text{charging}} = \Delta G_{i-\text{s}} \qquad (15.2)$$

If the electrostatic work problems can be solved, this model will provide results in terms of ΔG, which is advantageous, since it is already clear that by measuring enthalpies and calculating entropies, an empirical check of the model and its assumptions can be made. Thus, the stage is set for the solution to this problem.

The work represented by the first term, $w_{\text{discharge}}$, occurs in a vacuum and can be determined by solving the problem of charging a sphere in vacuum. This answer is found by bringing infinitesimal bits of charge from infinity to the sphere until the sphere has the correct charge. The solution to this problem is

$$\int dw = \int_0^{z_i e_o} \psi\, dq = w_{\text{charging}} = \frac{(z_i e_o)^2}{8\pi \varepsilon_o r_i} \tag{15.3}$$

(see Appendix O). Since the work of discharging is just the opposite of the work of charging

$$w_{\text{discharge}} = -\frac{(z_i e_o)^2}{8\pi \varepsilon_o r_i} \tag{15.4}$$

The work of charging the sphere in solvent is similar, except that the presence of the solvent, no matter what its structure, requires consideration of its dielectric constant, ε. The physical treatment of the alteration in the electric field has already been discussed and, in this problem of structureless solvent, is adequate for understanding what is to follow. A molecular–structural treatment of the dielectric constant will be forthcoming.

Earlier, it was shown that the electrostatic force can generally be written as

$$F = \frac{q_1 q_2}{4\pi \varepsilon_o \varepsilon r^2} \tag{15.5}$$

where in a vacuum ε is equal to 1. When a medium other than a vacuum exists between the two point charges, the force will be less depending on the medium's ability to modify the electrostatic force. In this case, the dielectric constant of the continuum (solvent) is chosen. The following equation for the work of recharging the ion in the solvent is obtained:

$$w_{\text{charging}} = \frac{(z_i e_o)^2}{8\pi \varepsilon_o \varepsilon r_i} \tag{15.6}$$

It is now possible to write the equation describing the free energy of the ion–solvent interaction:

$$\Delta G_{i-\text{s}} = w_{\text{discharge}} + w_{\text{charging}} \tag{15.7}$$

Substituting the terms from Eqs. (15.4) and (15.6) gives

$$\Delta G_{\text{i–s}} = -\frac{(z_i e_o)^2}{8\pi \varepsilon_o r_i} + \frac{(z_i e_o)^2}{8\pi \varepsilon_o \varepsilon r_i} \tag{15.8}$$

Rearranging algebraically gives

$$\Delta G_{\text{i–s}} = -\frac{(z_i e_o)^2}{8\pi \varepsilon_o r_i}\left(1 - \frac{1}{\varepsilon}\right) \tag{15.9}$$

This result is for a single ion. To make experimental analysis more practical, it is useful to describe ΔG on a per mole basis:

$$\Delta G_{\text{i–s}} = -N_{\text{A}}\frac{(z_{\text{i}}e_{\text{o}})^2}{8\pi\varepsilon_{\text{o}}r_{\text{i}}}\left(1-\frac{1}{\varepsilon}\right) \tag{15.10}$$

This expression defines the energy of interaction between the solvent and a mole of ions. The equation shows that the energy of interaction will depend on the charge on the ion and its radius alone, assuming that the dielectric constant does not change under the conditions of the experiment. Since our interest is to study a model of aqueous electrolyte solutions, the experimentally known dielectric constant for bulk water, 80, will be used for the calculation of the ion–solvent interactional energy in the Born model.

15.2.2 Choosing an Experimental Observable to Test the Model

The point of this exercise is to determine whether a model of the interaction of rigid charged spheres and a structureless continuum has any validity as a description for aqueous electrolyte behavior. The next step is to define a protocol that can be used to compare the theoretical behavior of the ions to empirical results derived from experiment. The most practical experimental approach is to measure the enthalpy of the ion–solvent interaction. ΔG is related to enthalpy through the following relationship:

$$\Delta G = \Delta H - T\Delta S \tag{15.11}$$

The enthalpy is

$$\Delta H_{\text{i}-\text{s}} = \Delta G_{\text{i}-\text{s}} + T\Delta S_{\text{i}-\text{s}} \tag{15.12}$$

Before the model can be compared to experiment, some expression for $\Delta S_{\text{i}-\text{s}}$ must be found. We know that

$$-\Delta S = \frac{\partial \Delta G}{\partial T} \tag{15.13}$$

Therefore, $\Delta S_{\text{i}-\text{s}}$ can be found by differentiating $\Delta G_{\text{i}-\text{s}}$ with respect to temperature (T). It can be shown experimentally that the dielectric constant varies with temperature, and therefore temperature must be treated as a variable in the differentiation

$$\Delta S_{\text{i–s}} = -\frac{\partial \Delta G_{\text{i–s}}}{\partial T} = N_{\text{A}}\frac{(z_{\text{i}}e_{\text{o}})^2}{8\pi\varepsilon_{\text{o}}r_{\text{i}}}\frac{1}{\varepsilon^2}\frac{\partial\varepsilon}{\partial T} \tag{15.14}$$

Then solving for $\Delta H_{\text{i}-\text{s}}$

$$\Delta H_{\text{i–s}} = -N_{\text{A}}\frac{(z_{\text{i}}e_{\text{o}})^2}{8\pi\varepsilon_{\text{o}}r_{\text{i}}}\left(1-\frac{1}{\varepsilon}-\frac{T}{\varepsilon^2}\frac{\partial\varepsilon}{\partial T}\right) \tag{15.15}$$

The solutions of this equation are a series of straight lines that depend on the magnitude of charge, z_i, of the ion and the ionic radius. An example of a solution is drawn in Fig. 15.3. Now we need to compare the theoretical values to experimental enthalpies of solvation. How to proceed? What in the empirical world corresponds to the radius of the rigid Born sphere? How can the enthalpy of solvation for a single ion be measured? The first question is relatively easily answered. Since this approach supposes that the ion can be represented as a rigid sphere, crystallographic radii of ions obtained from x-ray crystallography will be used. The second question requires a somewhat more complicated examination.

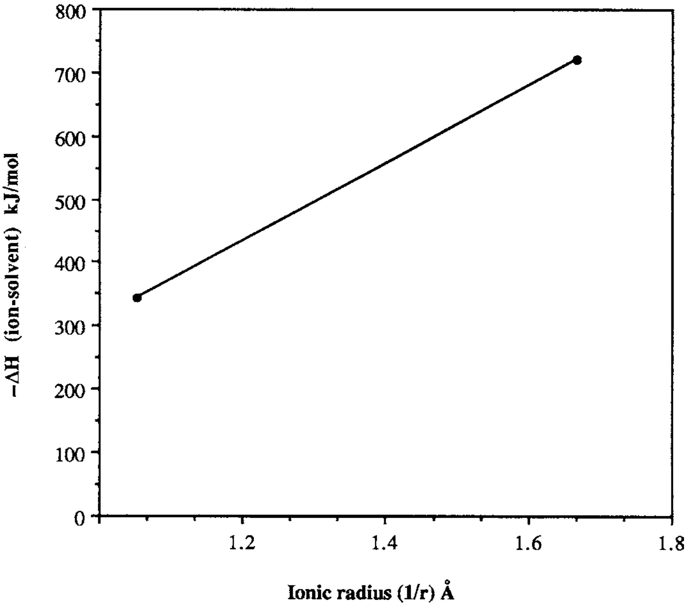


Fig. 15.3 ΔH of interaction for ions of charge z_x as calculated from the Born model

How can we make the required solution of a single ionic species in which the heat of solvation will be measured? Figure 15.4 illustrates this problem. Consider the events that would occur as this solution of single ions is made. Initially, the solvent has a neutral electrical charge. Now ions of a single type are added to the solution. As the ions are added, an unbalanced charge is added to the solution and an ever-growing coulombic force is generated. As each additional ion is brought to the solution, an increasing amount of energy is required to provide the work necessary to move the ions into solution. The difficulty with this scenario is that the energy measured in the production of this solution represents both the coulombic work of getting the ions to the solution and the energy of solvation. Such an approach will not generate the enthalpy number that is being sought.

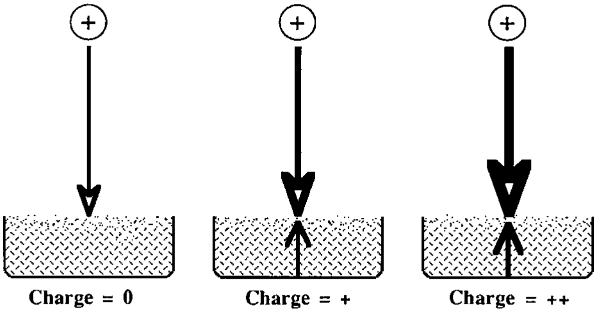


Fig. 15.4 Illustration of significant difficulties encountered while making a solution of one ionic species

Beyond the theoretical problems with this approach of adding a single unbalanced charged species to a solution, the usual practical approach to making an electrolytic solution is to dissolve a salt in the appropriate solvent. For example, if a NaCl crystal is being used, the positively charged sodium cations are added to an equal number of negatively charged chloride anions. This solution maintains electrical neutrality, so the enthalpy measured will represent the solvent–solute interaction and will contain no added term that represents work done in overpowering an electric force field. But while the generation of an electrolyte solution by the addition of a salt has eliminated the electroneutrality problem, the solution created now has two species of ions instead of the single species that has been modeled. This problem in obtaining information pertaining to the thermodynamics of a single ion in solution is an important one to solve because many questions in biological systems are concerned with the effects of single ions on a cellular system. Through the use of a relative scale of enthalpies, thermodynamic information about single ion enthalpies can be indirectly derived with reasonable confidence. This indirect method involves using the enthalpies of solvation from a series of salts, each of which shares a common ion, and attributing the differences in enthalpy to the single ion that is not commonly held. For example, the enthalpy of solvation for the single ion sodium could be determined if the enthalpies of solvation for NaCl and KCl were used. Methodology such as this therefore allows the enthalpies of solvation for a single ionic species to be considered. The relative differences found are generally meaningful only if they are linked to some relative point on a scale of other ions. This requires the assumption that the enthalpy of solvation of a salt can be found in which exactly one half of the enthalpy change can be assigned to the solvation of the cation and the other half to the solvation of the anion. Typically, KF is the salt chosen for this assumption because of the virtually identical crystallographic radii of the K^+ and F^- ions. If it appears that there is an element of circularity to this reasoning, that impression is correct. The problem with obtaining the experimental heats of solvation for single ions is that numerical assignment depends at some point on an assumption based on theoretical grounds.

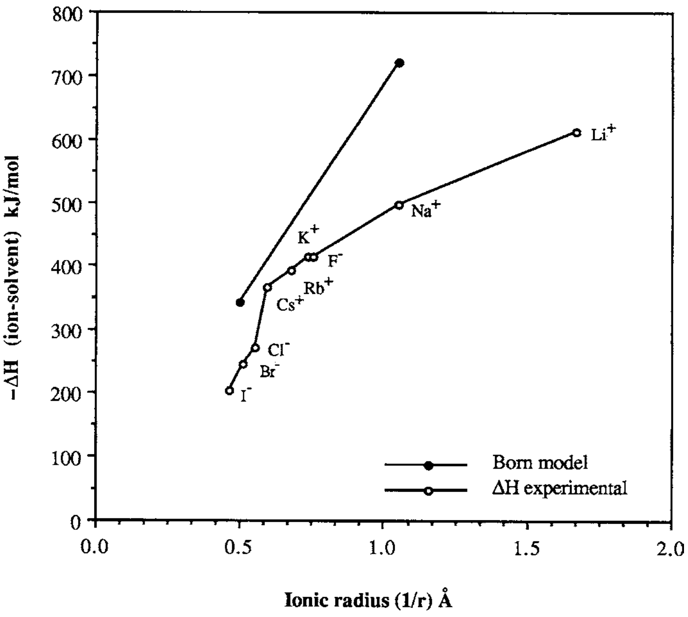


Fig. 15.5 Comparison between the experimental and theoretical values of the enthalpies of hydration for monovalent ions

Now comparison of experimental data with the Born model can be undertaken. In Fig. 15.5 the enthalpies of solution for a number of monovalent ions of varying crystallographic radii are compared to the Born solution for the rigid spheres of the same charge and the same radius. While trends between experiment and the model match, there are clearly major discrepancies between the model and the real world. Historically, with empirical corrections to the Born model greater agreement between the experimental and theoretical curves could be achieved. These corrections took two approaches. It can be shown that by simply adding 10 pm to the crystallographic radii of the positive ions and 85 pm to the negative ions, the experimental data can become nearly linear, in agreement with the model. It is also possible to consider that the use of the dielectric constant of 80 for the continuum may be incorrect. If, for example, there is an interaction between the charged sphere and the substance of the continuum, could the dielectric constant be changing? These corrections, of course, strike directly at the heart of the very assumptions on which the hypothesis was originally based. Both suggest that there is something discontinuous about the solvent and something wrong with simply using the crystallographic radii of ions for the size of the rigid sphere. Both suggest that there are structural considerations that must be included in an examination of the interaction between the ion and its aqueous solvent.

15.3 Adding Water Structure to the Solvent Continuum

In the last chapter, the structure of molecular water in its bulk form was presented. It is important to reemphasize several features. Water is a polar molecule with a dipole moment of 1.85 debye. The negative end of this dipole is oriented toward the oxygen atom and the positive end toward the hydrogen atoms. The molecule is ideal for forming hydrogen bonds, and in bulk water these bonds are extensively formed with adjacent water molecules. Because of this hydrogen bonding nature, ice, the crystalline form of water, forms an extensive lattice structure comprised of a puckered hexagonal array of water molecules. While liquid water does not have as highly ordered a lattice structure, significant portions of liquid water maintain the ordered tetrahedral clustering characteristic of ice. At any particular time, water molecules that are not part of the lattice are free to move in and out of the interstitial spaces formed by the structure. The local arrangement of water molecules is in constant flux as some free molecules reassociate with the lattice, and some lattice molecules break off to become free in the interstitial spaces. Many of the thermodynamic qualities of water depend on the hydrogen bonding nature and the structure of the bulk water. What is the effect on the water structure and hence on the properties of the solution, when ions (and later macromolecules) are added?

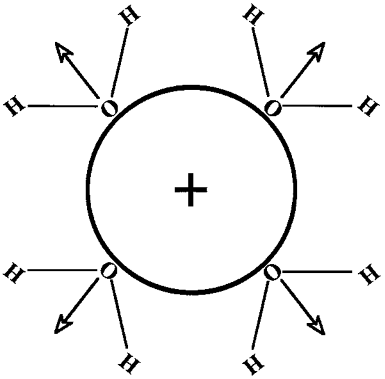


Fig. 15.6 The interaction of a charged ion with unassociated water molecules. There is an orientation of the water molecules, which behave as dipoles, because of an ion−dipole interaction

For simplicity's sake, consider what will happen if an ion is added to water molecules that are all free and not in any arrayed structure. Figure 15.6 illustrates this case. The ion is charged and is the center of an electrical force that emanates symmetrically from the ion. This electrical force will seek to orient any other electrically polar object near it in such fashion that like charges will be pushed away and unlike charges will be attracted. The free water molecules near the ion, being electrically polar, will be subject to this orienting force. If the ion is positively charged, the negative end of the water molecule will be turned toward the ion; if the ion is negative, the water will orient its positive end toward the ion. Now consider the case of structured liquid water. As the ion now exerts its reorienting force on the water molecules, the water molecules are subject to ion–dipole forces that may tear water molecules out of the lattice structure as it orients them toward the ionic point charge. What is different about this model with structured water as the solvent in comparison to the Born model? A principal difference is that the continuum model considered all ion–solvent interactions to be simply electrostatic in nature, while this structured water model considers the interaction to be ion–dipole in nature.

15.3.1 The Energy of Ion–Dipole Interactions Depends on Geometry

Understanding ion–dipole interactions requires a sense of the energy of the interaction, $U_{\text{i–d}}$. A qualitative understanding is satisfactory for this discussion. In a dipole, the separation of charge occurs over a finite distance. This characteristic is reflected in the definition of the dipole moment as being the product of the electric charge times the distance separating the two centers. The energy of interaction between ion and dipole will be the energy between the dipole center and the ion center. The line, r, connecting these two points makes an angle θ with the dipole (see Fig. 15.7). The energy of interaction is given by the product of the charge on the ion, $z_{\text{i}}e_{\text{o}}$, and the potential, ψ_r, i.e., $z_{\text{i}}e_{\text{o}}\psi_r$. Finding the potential, ψ_r, due to the dipole can be accomplished if the charges at each end of the dipole are considered separately and then added, according to the law of superposition of potentials. The distance between the charges at each end of the dipole and the ion center, however, is not equivalent to r but rather to two other radii, r_1 and r_2. Therefore, the potential can be written as

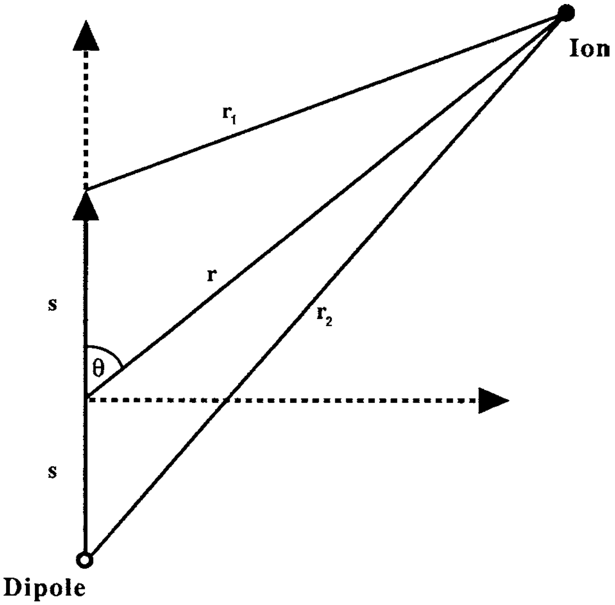


Fig. 15.7 Geometry for the calculation of $U_{\text{i–d}}$

$$\psi_r = \frac{q}{r_1} + \frac{-q}{r_2} \tag{15.16}$$

The distances r_1 and r_2 can be found by the Pythagorean theorem. The result has in it a term that includes the distance from each of the ends of the dipole to the midpoint of the dipole. If at this point a very important assumption is made that the distance separating the ends of the dipole from the midpoint is insignificant compared to the distance between the dipole center and the ion center, the following relationship can be ultimately derived:

$$U_{\mathrm{i\text{-}d}} = -\frac{z_{\mathrm{i}} e_0 \mu \cos\theta}{4\pi\varepsilon_0 \varepsilon r^2} \tag{15.17}$$

Ion–dipole interactions are relatively strong and will depend on the angle of the dipole to the ion, the dipole moment, the charge on the ion, and the distance separating the ion and the dipole. However, as the dipole approaches the ion more closely, the distance r will more closely approximate the ignored distance between the dipole center and the dipole ends. At close proximity, the simplifying assumption of this equation will become unreasonable.

15.3.2 Dipoles in an Electric Field: A Molecular Picture of the Dielectric Constants

Before going back to the question of water structure near an ion, consider the role of dipoles in an electric field. An ideal model system for this subject is the *parallel plate capacitor* (Fig. 15.8). Remember from our earlier discussion that the

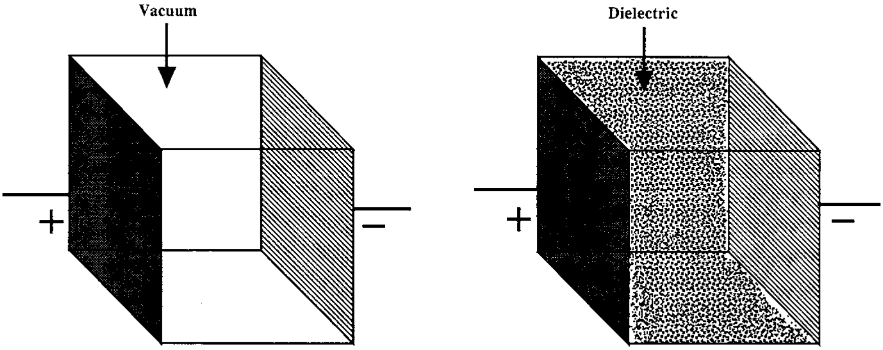


Fig. 15.8 The parallel plate capacitor can be used as a system to understand dipole behavior in an electric field. On the left, the plates are separated by a vacuum; on the right, the plates are separated by a dielectric material

mathematical description for an electrical field in parallel plate capacitor is straightforward. Previously we noted that capacitors store electrical energy as a stored charge. The charge and capacitance are related:

$$Q = C \times V \tag{15.18}$$

the terms being defined earlier. In a parallel plate capacitor, the capacitance is given by the equation

$$C = \frac{\varepsilon A}{d} \tag{15.19}$$

where A is the area of the plates, ε is the dielectric constant of the material between the plates, and d is the distance between the plates. Current will not flow in capacitors because the plates are separated by an insulating dielectric. When an external voltage is applied across the plates, charge is moved from the battery to the plates. As the plates charge, an electric force is ultimately generated equal in magnitude to the voltage of the external source. However, the field direction across the capacitor is opposite to that of the external source. Figure 15.9 illustrates the potential across the capacitor.

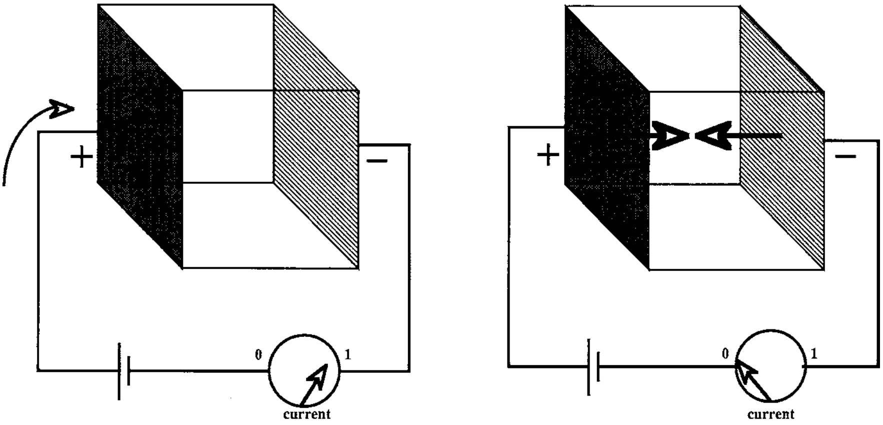


Fig. 15.9 The buildup of charge on the plates of the capacitor leads to a counter field that exactly opposes the potential of the external source

The number of fundamental charges necessary to generate the counterpotential is given by Eq. (15.18). At this point, the voltage or electric field strength across the capacitor is equal to the external force. The capacitor is now disconnected from the circuit. Since the charges on the plates have nowhere to go, the electric field measured across the capacitor will still be the same as the original field strength. If the dielectric is a vacuum, there are no molecules between the plates, and the counterpotential field extends through empty space. When a dielectric material is inserted

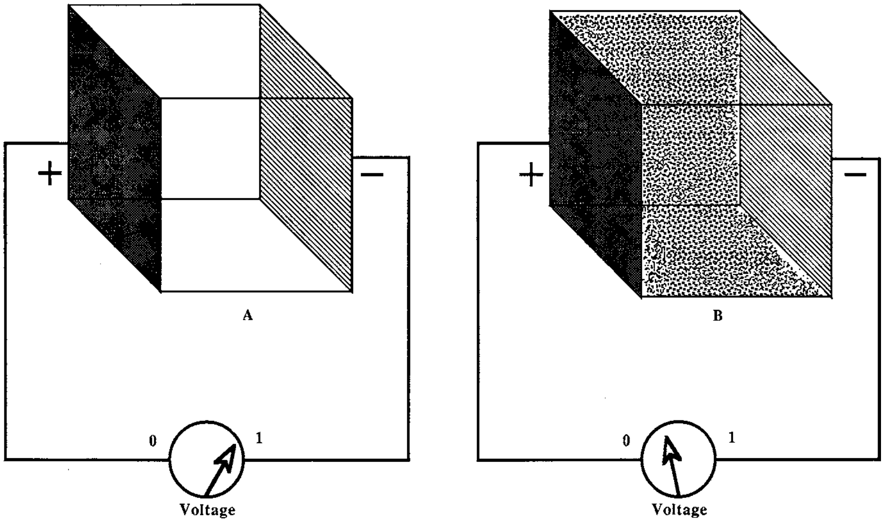


Fig. 15.10 The measured potential across a charged capacitor is altered by the dielectric constant of the dielectric material separating the two plates

between the charged plates, the measured potential field across the capacitor diminishes. The amount of field attenuation depends on the dielectric material chosen (Fig. 15.10). There is no loss of charge since the same charge is still on the plates as when the capacitor was initially charged. This means that the capacitance of the parallel plate capacitor has increased solely because the dielectric has changed.

The dielectric constant is derived from a comparison between the capacitance of a parallel plate capacitor using a vacuum as the dielectric material and that of the same capacitor with a different dielectric material replacing the vacuum. The relationship is written as

$$\varepsilon = \frac{C_{\text{any dielectric}}}{C_{\text{vacuum}}} \tag{15.20}$$

What is happening in the dielectric that is changing the measured electrical force across the capacitor? Two cases will be considered: the first will be when the dielectric is comprised of molecules that are permanent dipoles; the second, when the molecules do not have a dipole moment.

In the case where the dielectric is comprised of *permanent dipoles*, how will the dipoles be oriented prior to insertion into the field between the capacitor plates? Measurements on a macroscopic timescale (i.e., microseconds or longer) show that no single overall orientation of the molecules as a group can be ascertained; the dipoles are oriented randomly. This must be the case since these dipoles comprise a neutral substance that, while made up of polar molecules, is itself apolar. When the randomly arranged dielectric is now placed within the field of the capacitor, the dipoles experience the electrical force and attempt to align themselves with the field.

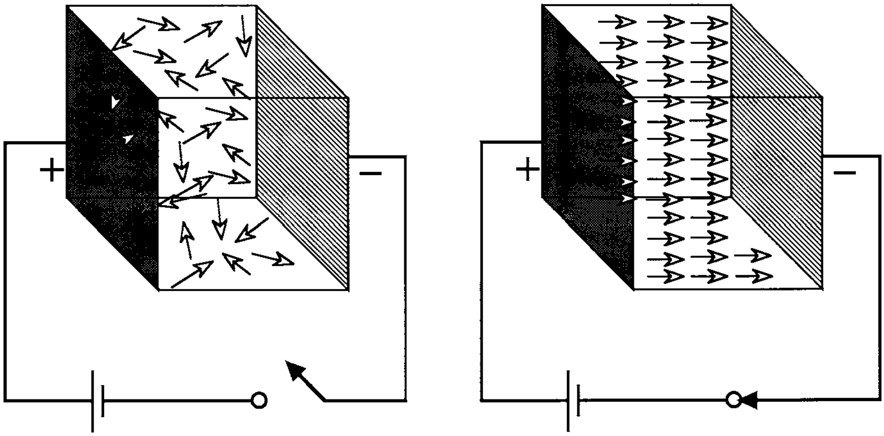


Fig. 15.11 Alignment of permanent dipoles in the absence and presence of an electric field. In the diagram on the left, there is a random arrangement of dipoles. When the field is switched on, the dipoles attempt to line up in the direction of the lines of electrical force

The positive end of the dipole will turn and align itself normal to the negative plate; conversely the negative end of the dipole will align itself normal to the positive plate (Fig. 15.11).

When a dielectric comprised of molecules without a dipole moment is placed into an electric field, the field will cause the displacement of the electron clouds (negatively charged) away from the nuclei (positively charged), thus inducing a dipole. Because these *induced dipoles* are generated by the field, they are aligned with the field at the instant of generation. They will exist as long as the field remains.

Thus an electric field has the ability to align permanent dipoles or induce dipoles along its lines of force at the molecular level. However, these aligned dipole charges are arranged with their charges opposite in orientation to the field. The separated charges of the dipoles themselves will generate an electric field, but one that is opposite to the field of the capacitor. This counterpotential can, by the principle of superposition, be shown to diminish the potential measured across the capacitor. It thus accounts for the increased capacitance of the system.

Are there other forces interacting between the dipoles and the electric field that should be considered? It will be worthwhile to investigate this question for it leads back to the original quest for an ion–dipole interaction in water solution. There are three further aspects that should be considered:

(1) First, what is the ease with which the permanent dipole can be aligned, and how susceptible is the non-permanent dipole to being induced into an aligned dipole? The measure of these properties is the polarizability of the material.
(2) The next consideration is, What are the effects of non-aligning forces on the orientation of the dipoles and on the overall average dipole moment of such dipoles? The primary considerations here are the randomizing thermal effects.

(3) Finally, the question of how the structure of the dielectric will respond to the various forces acting on it needs to be examined. Since the primary concern here is with water, the dielectric material of interest will be one containing permanent dipoles. Recognizing that, in the liquid state, lattice structural considerations will need to be considered, it will initially be simpler to consider each water molecule as free and independent (this corresponds of course to gaseous water, or steam).

The *susceptibility* of a material is the physical property that determines its dielectric constant. When a dielectric is placed in an electric field, there will be two forces acting against each other. The electric field caused by the excess charge, Q, on the capacitor plates either will attempt to align the permanent dipoles or will induce temporary dipoles. Depending on the ease with which these dipoles can be created or aligned, a charge separation, q_{dipole}, will occur. q_{dipole} has a field that is directed counter to the field resulting from the capacitor charge q. An analysis of this problem by Gauss' law gives the result that the electric force, X_{ext}, directed from one plate through the dielectric toward the second plate is

$$X_{\text{ext}} = 4\pi \left(q - q_{\text{dipole}}\right) \tag{15.21}$$

The q_{dipole} term depends on the number of dipoles that can align with the original capacitor-derived field. q_{dipole} is related to the ease of dipole alignment (α or the polarizability), the field that is causing the alignment (X_{ext}), and the number of dipoles aligned (n). n is dependent on both α and X_{ext}. The relation is given by

$$q_{\text{dipole}} = \alpha X_{\text{ext}} n \tag{15.22}$$

Combining Eq. (15.21) with Eq. (15.22) gives

$$X_{\text{ext}} = 4\pi q - 4\pi \alpha X_{\text{ext}} n \tag{15.23}$$

Ultimately, it can be derived that

$$\varepsilon - 1 = 4\pi \alpha n \tag{15.24}$$

This equation quantifies the idea that the dielectric constant is related to the ability of a material to respond to an electric field by generating a counterforce through dipole-based charge separation.

The dielectric constant depends on the number of *aligned* dipoles to generate the counter field. Therefore, the ease of dipole alignment, α, must be balanced against non-aligning or randomizing forces. In the earlier discussion of the Born model, it was noted that the dielectric constant varies with temperature. Thermal effects are a significant randomizing force in dielectrics. Imagine that as the dipoles swing into alignment with the external field, they are constantly buffeted by thermal collisions that randomly knock them out of alignment. The higher the temperature, the more likely is the chance of a significantly energetic thermal collision. Ultimately, a balance will be struck between the orienting force of the external field and the randomizing force of the thermal collisions. This balance can be written in a form of the *Boltzmann distribution law*:

$$n_{\mathrm{o}} = Re^{-w/kT} \tag{15.25}$$

Here, n_{o} represents the number of dipoles aligned at a particular angle θ to the external field, R is a proportionality constant, and w represents the work done by the molecule in orienting itself with the field:

$$w = -\mu X_{\mathrm{ext}} \cos\theta \tag{15.26}$$

The average orientation of the dipoles to the field will depend on the thermal energy in the system. As the system becomes hotter and hotter, the dipoles will have a greater tendency to be randomly oriented in spite of the aligning force of the external electric field. By using standard formulas for finding average values, a value for the average dipole moment depending on both field and temperature can be found:

$$\langle\mu\rangle = \frac{\mu^2 X_{\mathrm{ext}}}{3kT} \tag{15.27}$$

This formula gives the average dipole moment for a gas dipole. For a number of dipoles, n, the charge can be given by

$$q_{\mathrm{p}} = \frac{n\mu^2 X_{\mathrm{ext}}}{3kT} \tag{15.28}$$

Thermal effects have a greater disturbing effect on the permanent dipoles than on the induced dipoles. Molecules with induced dipole moments do not move into an orientation but rather are oriented as they are created by the electric field. The charge due to the dipole orientation, q_{dipole}, will be represented by a term containing permanent dipole contributions (q_{p}) and induced dipole contributions (q_{i}):

$$q_{\mathrm{dipole}} = q_{\mathrm{p}} + q_{\mathrm{i}} \tag{15.29}$$

Earlier, it was shown that the polarizability was related to q_{dipole}:

$$\alpha = \frac{q_{\mathrm{dipole}}}{nX_{\mathrm{ext}}} \tag{15.22}$$

Combining Eqs. (15.29) and (15.22) gives the following result:

$$\alpha = \frac{q_{\mathrm{p}}}{nX_{\mathrm{ext}}} + \frac{q_{\mathrm{i}}}{nX_{\mathrm{ext}}} \tag{15.30}$$

which is the same as

$$\alpha = \alpha_{\text{orient}} + \alpha_{\text{deform}} \tag{15.31}$$

The first term, α_{orient}, represents the contribution to the total polarizability of the dielectric molecules from the permanent dipoles. The second term, α_{deform}, represents the polarizability due to the ability of the electric field to deform the molecule and induce a dipole. This equation can be rewritten in terms of the dielectric constant:

$$\varepsilon - 1 = 4\pi n\alpha_{\text{orient}} + 4\pi n\alpha_{\text{deform}} \tag{15.32}$$

The first term can be rewritten in terms of Eq. (15.28), giving

$$\varepsilon - 1 = \frac{4\pi n\mu^2}{3kT} + 4\pi n\alpha_{\text{deform}} \tag{15.33}$$

When the temperature rises, the contribution to the dielectric constant from the permanent dipoles grows smaller and eventually becomes negligible, and the entire value of ε is accounted for by the induced dipoles in the system.

15.3.3 What Happens When the Dielectric Is Liquid Water?

We have only considered the interaction of individual dipoles and the orienting electric field. In other words, there is no structural constraint on these interactions. This scenario only holds for dipoles that exist in dilute gaseous form. How will the structure of liquid water affect the ability of an electric field to orient the dipoles of the water? We can answer the question by considering a single water molecule at the center of a tetrahedral array of water. As the electric field attempts to orient this particular dipole, it must orient the entire complex of water molecules. The issue is no longer simply one of the orientation of a permanent dipole countered by the forces of thermal disarray. Now the problem is aligning the entire array with the electric field. Attention must shift away from the average dipole moment of a single molecule to the average dipole moment of the ice-like structure. The average dipole moment of a tetrahedral cluster of water molecules is the vector sum of the dipoles for each element of the cluster:

$$\langle \mu_{\text{cluster}} \rangle = \mu + g\left(\mu\overline{\cos\gamma}\right) \tag{15.34}$$

where g, the number of nearest neighbors, is multiplied by the average of the cosine values between the central molecule's dipole moment (μ) and those of the neighbors. As can be seen, the dipole moment of the cluster will be greater than the electric moment of the single central dipole itself. A relationship between the dipole alignment and the thermal disorientation can also be made, and the following is the result:

$$\langle \mu_{\mathrm{group}} \rangle = \frac{\mu^2 \left(1 + g\overline{\cos\gamma}\right)^2}{3kT} X \tag{15.35}$$

where X is the electric field. (Proof that this is a general equation that includes the special case of the gaseous dipole derived earlier is left as an exercise for the reader.)

Is the electric force that originates at the capacitor plate the only force that is of consequence for the dipole residing inside the cluster of ice-like water? The interaction of the free dipoles in a dilute gas is limited because the individual dipoles are too far apart to feel the field emanating from the nearest dipole. In the case of a tetrahedral water cluster, the dipoles are atomically close to one another, and the electric field experienced by the central dipole will be a sum of the external electric field derived from the capacitor, X_{ext}, and also a more local field derived from the surrounding molecules, X_{loc}. Therefore, in the case of liquid water, the expression for the orientation polarizability (Eq. (15.22)) does not simply depend on the external field, as was the case for gaseous dipoles. The better expression for the field acting to orient the dipoles will be derived from both X_{ext} and X_{loc}.

It is possible to derive an expression for the electric field operating on a reference dipole or cluster of dipoles through the approach of Onsager. The reference dipole or grouping is considered to be surrounded by other molecules that, to a very reasonable approximation, can be thought of as forming a spherical cavity around the dipole grouping of interest. If the dipole cluster is removed, an empty cavity is the result, and the field inside the cavity (where the dipole would be located) is a resultant of both the external field and the field generated through the partial orientation of the surrounding dipoles. The expression for the local field inside the cavity acting on the dipole can be written as follows:

$$X_{\mathrm{loc}} = \frac{3\varepsilon}{2\varepsilon + 1} X_{\mathrm{ext}} \tag{15.36}$$

Since ε is always greater than 1, this equation says that the local field, and hence the orienting force on a dipole in a medium such as liquid water, is always greater inside the condensed phase than when the dipole is in a dilute gas phase.

What then is the effect on the dielectric constant of a condensed phase such as water, now taking into account the effects of structure and dipole interactions in this structure? This can be answered by starting with the earlier expression:

$$\varepsilon - 1 = 4\pi n\alpha \tag{15.24}$$

This equation was solved by determining α from the expression

$$\alpha = \frac{q_{\mathrm{dipole}}}{nX_{\mathrm{ext}}} \tag{15.22}$$

This was shown to be equal to

$$\alpha = \frac{q_{\mathrm{p}}}{nX_{\mathrm{ext}}} + \frac{q_{\mathrm{i}}}{nX_{\mathrm{ext}}} \tag{15.30}$$

Recall that the first term represents the orientation polarizability and the second the deformation polarizability. Earlier results, Eqs. (15.27) and (15.28), are combined to give the following expression for q_{p}:

$$q_{\mathrm{p}} = n\left\langle \mu_{\mathrm{group}} \right\rangle \tag{15.37}$$

Substituting the result for the average dipole moment of the group (Eq. (15.35)) and the local electric field X_{loc} (Eq. (15.36)) in place of the field X_{ext}, q_{p} becomes

$$q_{\mathrm{p}} = \frac{n\mu^2\left(1 + g\overline{\cos\gamma}\right)^2}{3kT}\frac{3\varepsilon}{2\varepsilon + 1}X_{\mathrm{ext}} \tag{15.38}$$

The q_{i} term may also be written as

$$q_{\mathrm{i}} = n\alpha_{\mathrm{deform}}\frac{3\varepsilon}{2\varepsilon + 1}X_{\mathrm{ext}} \tag{15.39}$$

When these substituted equations are used to evaluate Eq. (15.24), the following is the result:

$$\varepsilon - 1 = 4\pi n\frac{3\varepsilon}{2\varepsilon + 1}\left(\alpha_{\mathrm{deform}} + \frac{\mu^2\left(1 + g\overline{\cos\gamma}\right)^2}{3kT}\right) \tag{15.40}$$

This equation is the *Kirkwood equation* for finding the dielectric constant of a condensed medium. This equation takes into account the interactions that arise from the structured nature of the water molecules and also the effect of the localized field rather than the external field in calculating the dielectric constant of a medium such as water. It is clear that the dielectric constant of a medium such as water is profoundly affected by the cluster structure of the associated dipoles. The structural linking of dipoles will increase the number of nearest neighbors and hence increase the dielectric constant.

With the Kirkwood equation in hand, consider the case of liquid H_2S and H_2O (Table 15.1). H_2S has a dipole moment (1.02 debye) that is approximately one-half that of H_2O and, largely due to a lack of hydrogen bonding, does not have a strongly associated structure. The dielectric constant of water is approximately 10 times that of H_2S. While not all of the increased dielectric constant is attributable to the association of the dipoles in H_2O, the largest contribution is due to this structure and not to the increased dipole moment. To illustrate this point, consider a liquid that is

Table 15.1 Comparison of the dipole moments and dielectric constants of H_2O and H_2S, and $(CH_3)_2CO$

Substance	Dipole moment (debye)	Dielectric constant
H_2O	1.85	78.50
$(CH_3)_2CO$	2.90	20.00
H_2S	1.02	9.3

non-associating, yet whose molecules have a dipole moment even larger than that of water, $(CH_3)_2CO$. The dipole moment of $(CH_3)_2CO$ is over one and a half times larger than that of H_2O, yet the dielectric constant of $(CH_3)_2CO$ is four times less than that of H_2O. Thus, the Kirkwood equation predicts, and experimental data confirm, that it is the strong interactions between molecules in an associated structure that account for the high dielectric constant of water.

Finally, the Kirkwood equation indicates several parameters that are important for the definition of the dielectric constant. The most obvious is *g*, the number of nearest neighbors to the reference dipole. For liquid water, x-ray studies have indicated that approximately 4.5 water dipoles are coordinated around a reference central water molecule. Additionally, the Kirkwood equation explicitly gives the variation of dielectric constant with temperature. If this equation is used to calculate a dielectric constant for water at various temperatures, the theoretical values are usually within 10% of the experimental measurements.

15.4 Extending the Ion–Solvent Model Beyond the Born Model

By a circuitous route, we have now returned to the origin. What is the relationship between an ion and a structured solvent such as water? The ion will again be considered a charged sphere of fixed radius. There are several forces operating in this system, and the relationship of the water molecules to the ions will depend on the balance between these forces. Since the water molecules act as dipoles, there will be electrostatic forces acting between the ions and the dipoles. The strength of these electrostatic forces falls off according to the inverse square law and therefore is formidable near the ion but negligible at some distance from the ion. The hydrogen bonding interactive forces that will hold the water molecules together in ice-like clusters are present throughout the bulk of the water. Finally, there are thermal forces acting to disrupt the water structure. The ion–dipole interactions are relatively powerful forces, especially when compared to the hydrogen bonding that maintains the structure of the water. Near the ion, the electrostatic forces could therefore be expected to dominate, and water molecules will be oriented as dipoles to the ion's electrostatic field. These molecules will be torn from their association with other water molecules and will become immobilized or trapped around the ion. So tightly held are these water molecules that along with the ion they become a new kinetic body. The molecules in this solvent sheath are often referred to as *immobilized* or *irrotational* water, because of the virtually complete loss of freedom. At distances far removed from the ion, where the electrically orienting force is insignificant, the predominant forces will be those of the structure-stabilizing hydrogen bonding and the thermal effects that lead to a randomization of the structure. This is a description of the forces acting in bulk water, and indeed there is essentially no effect of the ion's electric field at these distances.

The two extremes are not difficult to comprehend, but what is the structure in the region where there is a balance between the orienting forces of the ion and the structure-forming forces of bulk water? In this region of balanced forces, the water molecules will be sometimes oriented to the ion-derived forces and at other times oriented to the structure of the bulk water. From a structural viewpoint, therefore, the time-averaged positions of water molecules in this region will give the appearance of a broken-down or less ice-like arrangement than that of bulk water. The water molecules in this region will be less associated and therefore have different physical properties, such as a decreased dielectric constant, when compared to bulk water. On the other hand, the ionic forces acting to orient and fix the water molecules to the ion will not be adequate to make the water in this region a permanent part of the same kinetic entity that the ion and the irrotational water sheath comprise. Consequently, while the irrotational water will move in lockstep with the ion when the ion moves, the water in this secondary sheath will not travel with the ion and its primary hydration sheath.

Solvent interactions with an ion therefore may be considered to occur in three layers (Fig. 15.12). First, there is a primary hydration sheath comprised of irrotational solvent molecules that move as the ion itself moves. As the distance from the ion increases, a secondary hydration sheath is entered where there is a partial ionic and solvent structural influence. Finally, the third layer is the bulk solvent itself and essentially feels no local force related to the ion.

15.4.1 Recalculating the New Model

When the idea of these hydration sheaths is incorporated into an analysis like the Born model, the predicted enthalpies of solution more closely fit the experimental values than those predicted by the Born equation. Such an approach was developed by Bernal and Fowler and by Elay and Evans. In a fashion analogous to the Born model, the heat of solvation, $\Delta G_{\text{i–s}}$, can be calculated by taking a rigid sphere representing an ion in a vacuum and bringing it into the solvent. Since the final structure in the solvent will be an ion surrounded by irrotational water and a secondary hydration sheath, the energy for the events leading to this structure must also be considered. The energy calculations will include enthalpic terms related to both the Born charging of the ion and the energy needed to disrupt the water structure and rearrange the solvent in the solvation sheaths. Furthermore, there will be entropic changes resulting from the alterations in degrees of freedom as the water molecules experience altered structural environments. Because our analysis is only semiquantitative, we will remain focused on the enthalpic changes and ignore the entropic changes. In a similar fashion to the treatment of the Born model, a thermodynamic cycle can be constructed to derive an expression for $\Delta G_{\text{ion–solvent}}$.

How does consideration of water structure modify the Born model? A number of changes derive from explicit consideration of the structure of the solvent. Explicit consideration of the structure of the solvent will dictate treatment of the ion as a unit comprised of the ion and its irrotational sheath of hydration water rather than

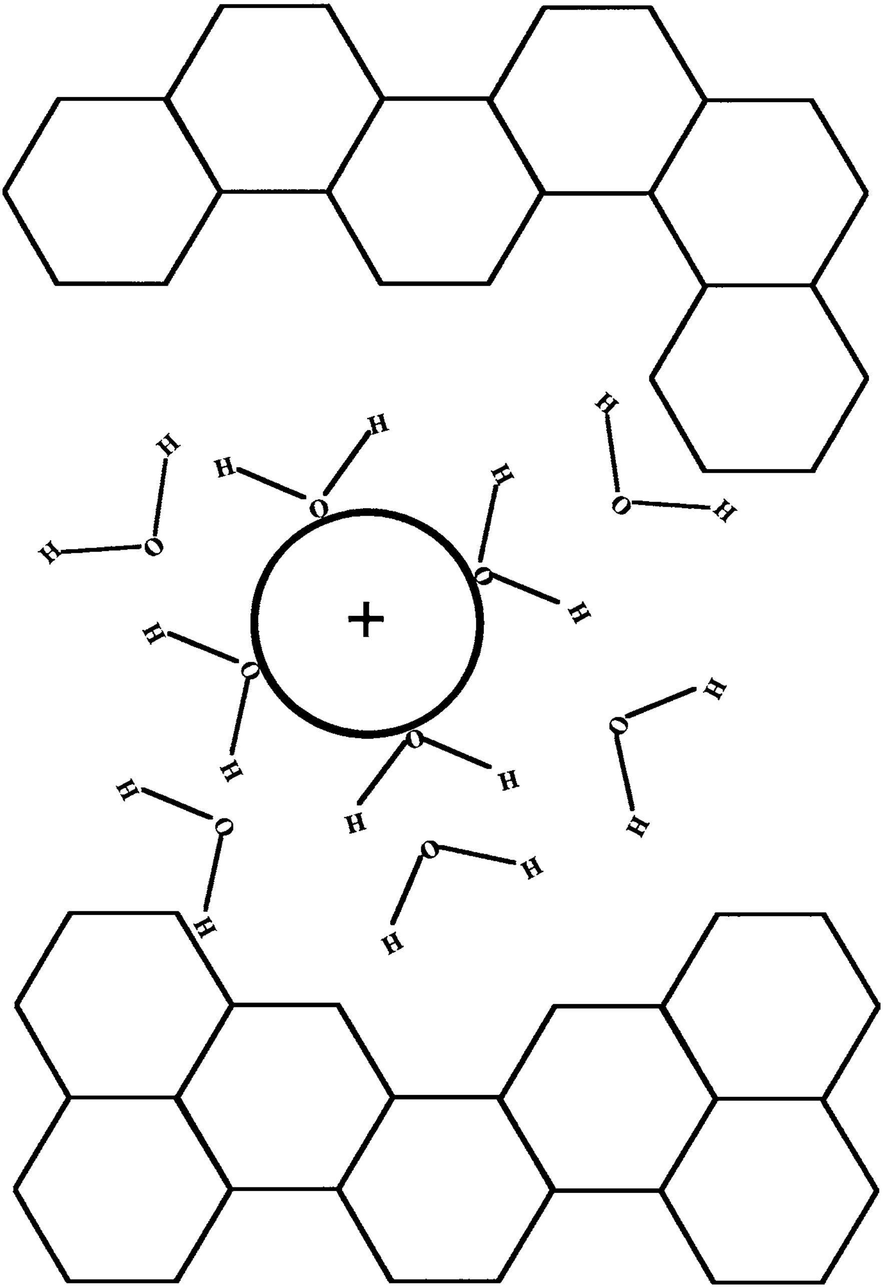


Fig. 15.12 An ion in water will be surrounded by an irrotational sheath, an intermediate sheath, and finally water with bulk properties

as a simple rigid sphere. Several new terms must be considered as a result of these changes. What are the changes in the new model?

(1) Rather than the crystallographic radius, the radius of the ion plus its hydration sheath will be used.
(2) The formation of the hydration sheath around the ion will require a specific number of molecules of water, n.
(3) The solvent has structure, and therefore the space occupied by the hydrated ion added into the solvent must be explicitly considered. Addition of the hydrated ion will require the removal of enough water molecules to leave space for the hydrated ion. The size of an ion is about the same as that of a water molecule, so the volume required for the hydrated ion will be related to the expression $V(n+1)$. In this expression, V is a function of the number of water molecules in the hydration sheath plus one volume (equivalent to a water molecule) for the space occupied by the ion.
(4) $n + 1$ water molecules will be removed from the solvent which will both make room for the hydrated ion and provide the molecules necessary to hydrate the ion (which starts in a vacuum). However, for every hydrated ion, one extra water molecule is taken out of the solvent. At some point in the cycle, these molecules must be returned to the solvent.
(5) Additional terms must also be added to account for the formation of the hydration sheath around the ion. These terms must include an amount of work required to dissociate the $n+1$ water molecules and the work of the ion–dipole interactions necessary to form n molecules into a primary hydration sheath.
(6) Finally, the secondary hydration sheath will need to be formed after the ion has been moved into the solvent.

Once all of the terms listed above are considered, then the problem can be treated in a fashion similar to the earlier Born model. Figure 15.13 outlines the terms of the cycle described here.

The enthalpy of ion–solvent interaction for this model can be written as the sum of the various terms just discussed. Table 15.2 provides a listing of these terms. As Table 15.2 indicates, the heat of interaction will depend on a term reflecting a process similar to the Born charging process derived earlier and also on a term that describes the ion–dipole interactions resulting from the formation of the primary solvation sheath. The amount of work necessary to remove n water molecules, dissociate them, and condense the extra water molecules as well as the energy of interaction of the primary solvated ion with the secondary solvation sheath must also be known. It turns out that, on a molar basis, this work function is 84 kJ/mol for positive ions and 126 kJ/mol for negative ions. These numbers are derived from a consideration of the breaking and re-formation of the hydrogen bonds in the water structure as the water molecules are removed, dissociated, replaced, and reoriented in the cycle. If the number of hydrogen bonds broken and then ultimately re-formed is added up and multiplied by the energy of hydrogen bond formation (–21 kJ/mol), the values will be found for a tetrahedral water structure. The difference in energy

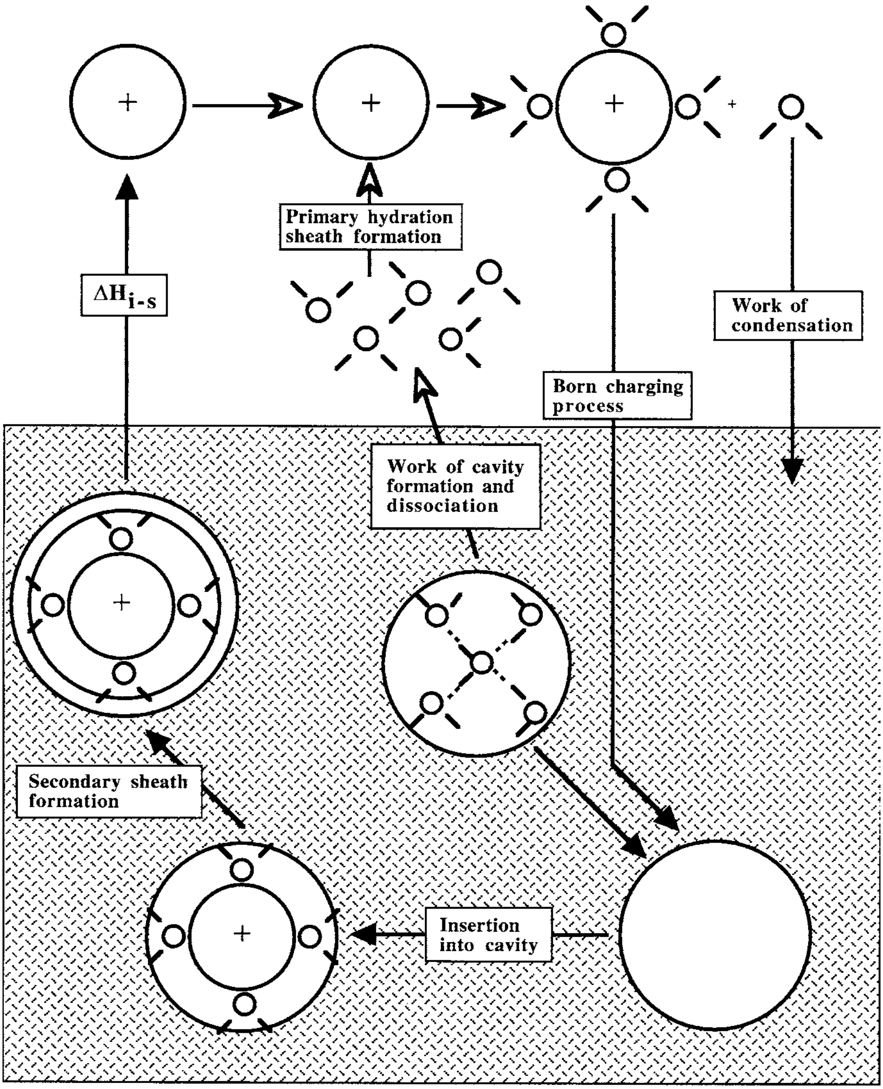

+
Primary hydration sheath formation
ΔH$_{i-s}$
Work of condensation
Born charging process
Work of cavity formation and dissociation
Secondary sheath formation
Insertion into cavity

Fig. 15.13 Schematic of thermodynamic cycle for the refinement of the Born model

between the positive and negative ion interaction occurs because of the different orientation of the water molecules in the primary sheath. In the case of the positive ion, the negative pole of the water dipole is closest to the ion; and for the negative ion, the positive pole of the dipole orients to the ion. In the case of the positive ion, there are two fewer hydrogen bonds re-formed after the water-removal–dissociation–reassociation process, while four less hydrogen bonds are re-formed in the case of the negative ion. The structural proof of this statement is left as a problem for the reader. The work required for the return of one mole of water for each mole of ion added to the solvent is the *latent heat of condensation*, about 42 kJ/mol. By adding together the heat of condensation and the energy necessary to break the requisite number of hydrogen bonds in each case, the numbers given in Table 15.2 are found.

Table 15.2 Components for determination of $\Delta H_{\text{ion–solvent}}$ for the extended Born-type model

Born charging[a]	Ion−dipole interaction[b]	Hydrogen bond/ condensation (kJ/mol)
$-\frac{N_A(z_ie_o)}{8\pi\varepsilon_o(r_i+2r_i)}^2\left[1-\frac{1}{\varepsilon}-\frac{T}{\varepsilon^2}\frac{\partial\varepsilon}{\partial T}\right]$	$\frac{N_A n z_i e_o \mu}{4\pi\varepsilon_o(r_i+r_s)^2}$	126[c] 84[d]

[a] Note that the radius has been changed to reflect the radius of the ion plus its hydration sheath

[b] Note that the dielectric constant does not appear in this equation because the orientation of the dipoles into the primary sheath is considered to occur in a vacuum

[c] For negative ions with a hydration sheath of four water molecules

[d] For positive ions with a hydration sheath of four water molecules

These considerations allow the following equations to be described for the ion–solvent interaction in an aqueous solvent, where water is considered a dipole and the coordination number of the water in the primary sheath is four:

For negative ions:

$$\Delta H_{\text{i–s}} = 126 - \frac{4N_A z_i e_o \mu_W}{4\pi\varepsilon_o\,(r_i + r_W)^2} - N_A\frac{(z_ie_o)^2}{8\pi\varepsilon_o\,(r_i+2r_W)}\left(1-\frac{1}{\varepsilon}-\frac{T}{\varepsilon^2}\frac{\partial\varepsilon}{\partial T}\right) \tag{15.41}$$

For positive ions:

$$\Delta H_{\text{i–s}} = 84 - \frac{4N_A z_i e_o \mu_W}{4\pi\varepsilon_o\,(r_i + r_W)^2} - N_A\frac{(z_ie_o)^2}{8\pi\varepsilon_o\,(r_i+2r_W)}\left(1-\frac{1}{\varepsilon}-\frac{T}{\varepsilon^2}\frac{\partial\varepsilon}{\partial T}\right) \tag{15.42}$$

where r_W and μ_W represent the radius of the hydration sheath in water and the dipole moment of water, respectively.

Do these added terms increase the accuracy of the theoretical model in its predictions of the heats of hydration of individual ions? Figure 15.14 shows the calculated versus the empirical heats of solvation for the ions. Consideration of the structural aspects of ion–solvent interactions leads to a vast improvement in the agreement between model and experiment.

This model suggests that the major contributions to the energy of interaction between solvent and ions come from the electrostatic terms of the modified Born charging model and the dipole–ion interactions of the solvation sheaths. It can be shown in a more rigorous analysis that a few more modifications can lead to an even better model. While such an analysis is beyond our intended scope, it is to be

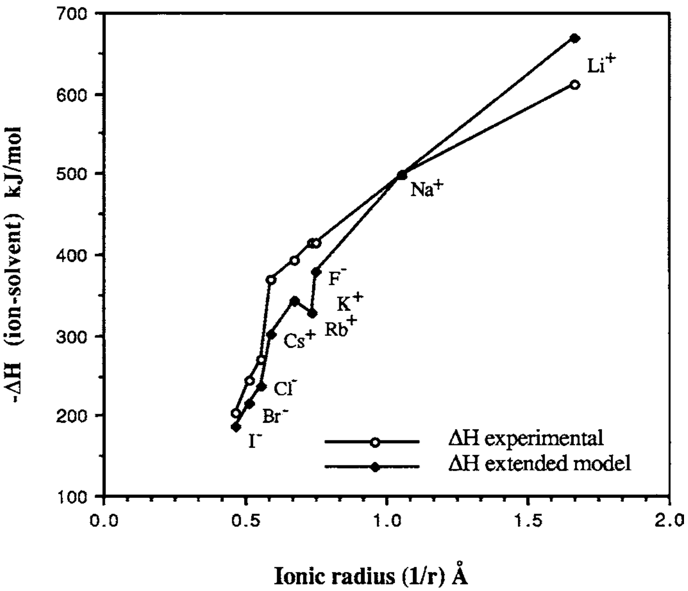


Fig. 15.14 Comparison of $\Delta H_{\text{ion–solvent}}$ from experiment and extended Born theory

noted that when water is considered as a quadrupole (an assembly of four charges), as originally described by the Bjerrum model of the molecule (cf. Chapter 20), the energetics of the interactions in the primary solvation sheath are more accurately described. The most important concept to understand is that the nature of the ion–solvent interaction in aqueous solutions is highly dependent on the polar nature of the water molecules.

15.5 Solutions of Inorganic Ions

In summary, small ions disrupt the organized structure of liquid water. The electrostatic interaction between the ion and the water dipole forces an orientation of the water molecules with respect to the ion. The hydrogen bonds between water molecules are weakened and then broken, as they deviate from their normal 180° during their reorientation. In the final fully hydrated structure, each ion is surrounded by water molecules oriented such that the partial opposite charge faces the ion. The number of water molecules involved in the primary hydration sheath depends upon the ion's size and charge.

The enthalpy associated with the hydration of these ions is always negative ($\Delta H < 0$), with multivalent ions having a much more negative ΔH than monovalent ones (Table 15.3). The enthalpy of solvation is sizable and is attributable to a combination of interactions in the primary and secondary hydration sheaths, as well as the hydrogen bond energy lost when the bulk water structure is perturbed.

Table 15.3 Enthalpies of hydration for single ions

Ion	$\Delta H_{\text{hydration}(\text{kJ/mol})}$
H^+	−1071
Li^+	−548
Na^+	−485
K^+	−385
Ca^{2+}	−1715
Al^{3+}	−4807

Additionally, there will be an entropy decrease, its magnitude depending on the number of H_2O molecules involved. This decrease in entropy is a consequence of the removal of H_2O molecules from the bulk solvent, where the entropy is high, to the lower entropic environment of the hydration sheaths. The entropy falls because the number of possible microstates in bulk water is much higher than the number of relatively fixed locations around an ion. The enthalpy and the entropy of aqueous solvation of ions depend partly on the identity of the ion and not solely on its charge.

The existence of an ionic atmosphere in which normal hydrogen bonds do not exist accounts for the fact that the heat capacity of a dilute solution of an inorganic electrolyte (such as KBr) is very much smaller than that of water and decreases as the concentration of electrolyte increases (Fig. 15.15). Furthermore, because the hydrogen bonds do not exist near the ion, the hexagonal structure with its large spaces is disturbed. This leads to a smaller volume of a molal solution of an inorganic salt when compared to the volume of pure water alone (i.e., $\Delta V < 0$).

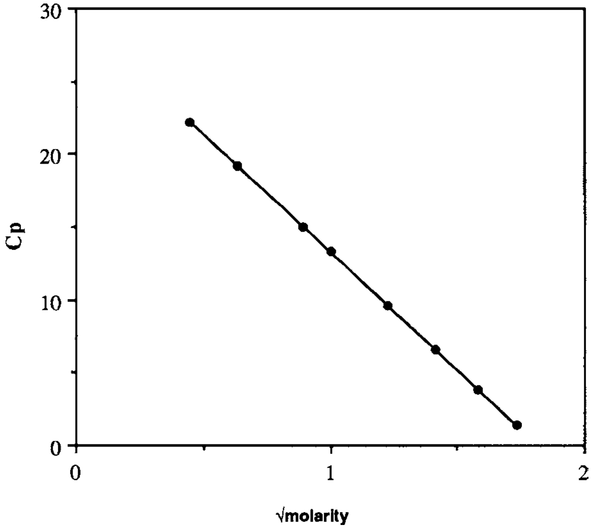


Fig. 15.15 Plot of the heat capacity (C_p) for a solution of KBr in water at various concentrations

The discussion here has centered on the solvation of small inorganic ions and on the effect of this solvation sheath on thermodynamic parameters such as heat capacity, enthalpy, and entropy of solution. In spite of the increased level of organization of the H_2O molecules nearest to an inorganic ion, the overall effect of solvation of these ions in water is to lower the energy requirement for any changes in the thermodynamic state of the system (e.g., heat capacity and heat of vaporization) in which disruption or formation of hydrogen bonding is involved.

15.6 Ion–Solvent Interactions in Biological Systems

At this point, it is clear that the interaction of an ionic species with its solvent will depend greatly on the polar and structural behavior of the solvent and on the electrostatic nature of the ion. Just as the behavior of the ion will be changed as a kinetic entity by the association of solvated molecules, the properties of the solvent will also be affected. This can be extremely important when the case of biological solutions is considered. Biologically relevant solutions contain mixtures of small polar and nonpolar molecules as well as macromolecules. A detailed treatment of the interactions of such molecules with water is forthcoming in the following chapters. Even without a detailed knowledge of the forces acting between non-ionic molecules and an aqueous solvent, it is pertinent to ask at this point what the effect of adding ionic species to an aqueous solution of other molecules will be.

The most obvious effect of the addition of ions to a solution will be manifest on the solubility of a particular molecular species. Consider the case of a macromolecule that has a limited solubility in water. The molecule is in solution because of its interactions with the water. If a quantity of ions is now added to the solution, the ions will usually strongly associate with the water, forming the familiar hydration sheaths. The association of the water molecules with the ion effectively removes water from the solution. Thus, the effective concentration of the water (its activity) decreases with the addition of the ions. If the ion that is added has a solvation number of four molecules of water, then for every mole of ion added, the effective concentration of water will be decreased by four moles. For example, in a liter of water that contains 55.5 mol of water, the addition of one mole of ions will leave only 51.5 mol of water free. This is a significant decrease in activity. The macromolecule that was barely soluble in water before has now been stripped of some of the water molecules necessary to keep it in solution, and it will precipitate out. This is one basic reason why the time-honored “salting out” step is so useful in protein purifications.

The reader should at this point have a much greater appreciation for the effect of the structure of water on the interactions between ionized solutes and water. So far, the only interactions that have been considered are those between the fields of the ion and the dipoles or quadrupoles of water molecules. The idea that an ion might interact with other ions has been ignored to this point. Few solutions are so dilute that such interactions can reasonably be ignored. Therefore, we now consider the more typical case of a solution where an ion will interact both with its solvent and with other ions as well.

Further Reading

General Texts

Bockris J. O'M., Reddy A.K.N., and Gamboa-Aldeco M.E. (1998) *Modern Electrochemistry*, edition. Kluwer Academic/Plenum, New York. (This revised three volume set is a prize and is required reading for any scientist interested in a lucid and complete grounding in the fundamentals of electrochemistry.)

Koryta J. and Dvorak J. (1993) *Principles of Electrochemistry*, edition. Wiley, New York.

Electrolyte Solution Structure

Arun Yethiraj A. and Chwen-Yang Shew C.Y. (1996) Structure of polyelectrolyte solutions. *Phys. Rev. Lett.*, **77**:3937–3940.

Bashford D. and Case D.A. (2000) Generalized Born models of macromolecular solvation effects. *Annu. Rev. Phys. Chem.*, **51**:129–152.

Hitoshi O.H. (1993) Structure and dynamics of hydrated ions. *Chem. Rev.*, **93**:1157–1204.

Im W., Michael F.M., and Brooks III C.L. (2003) An implicit membrane generalized born theory for the study of structure, stability, and interactions of membrane proteins. *Biophys. J.*, **85**:2900–2918.

Yanjie Z.Y. and Cremer P.S. (2006) Interactions between macromolecules and ions: the Hofmeister series. *Curr. Opin. Chem. Biol.*, **10**:658–663.

Problem Sets

1. Using the Born model, predict what the enthalpy of hydration for Na^+, K^+, Cl^-, $Ca2^+$, and OH^- will be in

   a. water at 368 K.
   b. CH_3CH_2OH at 298 K.
   c. dioxane at 298 K.

2. Use a table of the enthalpies of solvation for salts to derive a table of heats of solvation for single ions. Initially, use KF as the anion–cation pair that share the heat of solvation equally as the basis for the table. Then use another salt such as NaI. Try CsF. How does changing the pair of ions assumed to be equivalent affect the enthalpies of solvation for single ions? Predict what errors might be introduced in the empirical tests of the theories presented if the enthalpies of solvation of K+ and F– are not exactly equivalent as assumed.
3. Estimate the alteration in the dielectric constant in water that has formed pentagonal rather than hexagonal arrays around nonpolar molecules.

4. Show that the expression

$$\langle \mu_{\text{group}} \rangle = \frac{\mu^2 (1 + g \cos g)^2}{3kT} X$$

is a general equation that includes the special case of the electric moment of a gaseous dipole.

5. The enthalpies of hydration for single ions are substantially negative whether anions or cations are dissolved, yet the enthalpies of hydration of salts comprised of these ions are usually quite small and often slightly positive. Is this contradictory? How would you account for these observations?

6. Derive expressions for the enthalpy of ion–solvent interaction for positive and negative ions when the coordination number of the solvent sheath is six water molecules.

Chapter 16
Ion–Ion Interactions

Contents

16.1 Ion–Ion Interactions Can Be Modeled and These Models Can Be Experimentally Validated and Refined

If a potential electrolyte is added to an aqueous solution, it is conceivable that its dissociation may be so low that few ions are produced. It is also possible to make a solution from a true electrolyte so dilute that the number of ions in solution is very small. In both of these cases, the analysis completed in the previous chapter would be adequate to describe the interactions that cause the solution to deviate from ideality. These solutions would be so dilute that each and every ion in solution could, in a sense, look out past even its secondary hydration sheath and see only bulk water. These ions would thus seem to exist completely alone and isolated in solution.

In biological systems, however, ions are not often found in dilute conditions. Biological systems contain a great number of ions in solution even when the pro-

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_16, © Springer Science+Business Media, LLC 2010

teins, carbohydrates, and colloidal particles are ignored. What is the effect of other ions on a reference ion that looks out and feels the presence of other ions? The question is answered by finding out what forces act on the reference ion. Once these forces are identified, quantification of these forces can be used to study how the behavior of the reference ion will change. The test of the model will be to see how effectively the activity of the ion can be predicted.

It is worth taking a moment to reflect on how the understanding of these various interactions in an ionic solution is of value to the biological scientist. The aim of these reflections is twofold. First, it is desirable to arrive at an intuitive sense of how molecules will behave in the complex systems that characterize the biological world. Second, it is important to have some quantitative formulations that will allow these complex systems to be predicted and understood. When the biochemist or biophysicist measures a chemical change or a force acting in a biological system, the behavior of the reference ion, molecule, or force will be affected by the interactions throughout the system. Since the reference object is subject to the complex array of forces that are now being derived, it would be impossible to understand the implications of the measured quantity if the parts of the system are not appreciated.

If the interactions between a single ionic species and all the other species in solution are to be accurately described, it will be necessary to quantify the forces of interaction. What is sought is a description of the free energy of ion–ion interactions, $\Delta G_{i\text{–}i}$, in a system where the initial state is one of ions in solution with no ion–ion interaction, and the final state is one in which the ions in solution are interacting. The energy of this change in state is $\Delta G_{i\text{–}i}$. In most cases, and especially in cellular systems, the interest will be to describe the partial free energy change associated with a single ionic species with respect to the entire ionic assembly described by $\Delta G_{i\text{–}i}$. The required partial free energy change is given by the chemical potential, μ. The quantity $\Delta G_{i\text{–}i}$ will be given by the sum of the chemical potentials, $\mu_{i\text{–}I}$, for each ion that interacts with the assembly of other ions:

$$G_{i-i} = \sum n_i \mu_{i-I} \tag{16.1}$$

It is reasonable to assume (at least initially), as was done in the model of ion–solvent interactions, that the ion behaves as a charged rigid sphere and that the forces acting on it are essentially electrostatic in nature. Consequently, the chemical potential for a single ionic species will be related to the work of charging up a mole of the ions of interest while in the proximity of other ions. This is similar to the energy of interaction found in the Born charging process and can be written as

$$w_{\text{charging}} = \frac{(z_i e_{\text{o}})^2}{8\pi\varepsilon_{\text{o}}\varepsilon r_i} \tag{14.6}$$

In terms of the chemical potential, $\mu_{i\text{–}I}$, this can be written as

$$\Delta\mu_{i-I} = N_A w_{\text{charging}} \tag{16.2}$$

Combining these two equations and recognizing that $z_i e_o / 4\pi\varepsilon_o \varepsilon r_i$ is actually ψ, the electrostatic potential of the ion, the following can be written:

$$\Delta\mu_{i-I} = N_A \frac{z_i e_o}{2} \psi \tag{16.3}$$

The chemical potential change, then, of the interaction between the total ionic assembly and the ionic species of interest can be found by determining the electrostatic field at each individual ion that is a result of the other ions in solution. This field could be found if the spatial distribution of the ions in the solution were known relative to the reference ion. If such structural information were known, then the field could be calculated by the law of superposition, and the energy of interaction

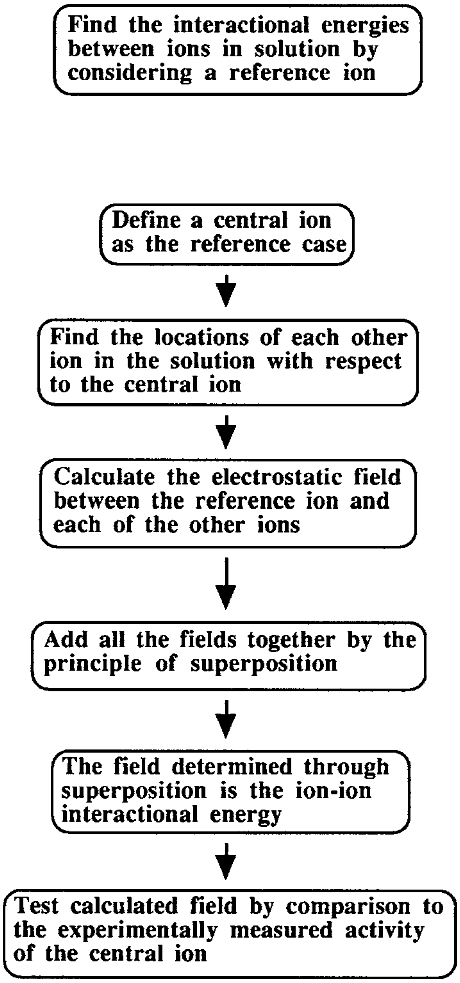

Find the interactional energies between ions in solution by considering a reference ion
Define a central ion as the reference case
Find the locations of each other ion in the solution with respect to the central ion
Calculate the electrostatic field between the reference ion and each of the other ions
Add all the fields together by the principle of superposition
The field determined through superposition is the ion-ion interactional energy
Test calculated field by comparison to the experimentally measured activity of the central ion

Fig. 16.1 General approach to solving the problem of ion–ion interaction forces

would result. Consequently, by constructing a model of the orientation of the ions that surround the reference ion and comparing the calculated energy of interaction with that found by experiment (by measuring activity), a test of the accuracy of the proposed structure will be possible. Figure 16.1 summarizes the approach.

16.2 The Debye–Hückel Model Is a Continuum Model That Relates a Distribution of Nearby Ions to a Central Reference Ion

The first step in solving this problem is to examine the distribution of charges that surround a reference ion as described by Debye and Hückel in 1923. In this analysis, a central or reference ion is considered and is given a specific charge. Like the Born model, the Debye–Hückel treatment assumes that everything other than this central ion is to be treated as a non-structured continuum of charge residing in a dielectric continuum. For the purposes of the discussion, the dielectric constant is taken to be that of bulk water. Now, consider the system just proposed. The central discrete ion is charged, and therefore, in the region close to the ion an electrical imbalance exists due to a charge separation. As pointed out previously, nature does not like regions of imbalance, and therefore an attempt will be made to neutralize this unbalanced charge distribution. The charge from the central ion will be neutralized by the continuum of charge that surrounds it. At equilibrium, the charge on the ion will be exactly countered by a countercharged atmosphere that will be arranged in some charge distribution around the ion as shown in Fig. 16.2. Locally, there will

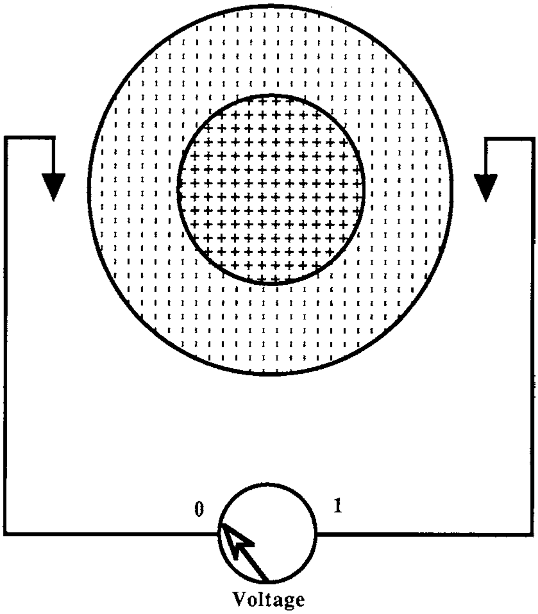


Fig. 16.2 The principle of electroneutrality gives rise to the existence of clouds of countercharge to balance the electric field emanating from the central ion

be regions of excess charge density, but, taken as a whole, the solution will be electroneutral since each of the central ions will be surrounded by atmospheres of charge that are exactly equal in magnitude but opposite in sign to the charge on the central ion. This is an important principle and is called the *principle of electroneutrality*. Mathematically, the principle can be expressed as follows:

$$\sum z_1 e_o X_i = 0 \tag{16.4}$$

where z_i is the number of elementary charges, e_0, carried by each mole fraction, X_i, of the species making up the entire solution.

Since the arrangement of the ions in the Debye–Hückel model is represented by the arrangement of a smeared charged atmosphere, describing the excess charge density around the central ion will provide information related to the spatial configuration of the actual ions in solution. In the earlier discussion of the dielectric constant, the balance between the directed electrostatic forces and the randomizing thermal forces was presented. As will shortly be seen, similar effects will need to be taken into account here.

It is important to make a historical note as to the origin of the idea of treating the interionic forces mathematically. Milner in 1912 first attempted to mathematically link the interionic forces acting in solution to the behavior of the solution. His treatment, however, was based on statistical mechanics; it was difficult to understand and therefore not easy to subject to experimental test. Consequently, the work went essentially unrecognized. The idea of the ionic atmosphere was proposed by Gouy in 1910. Gouy used the approach of smoothing out the ionic charges into a continuum and applying Poisson's equation. It was by building on these pioneering works that Debye and Hückel made their own highly significant contributions that led to an experimentally testable model.

The model of Debye and Hückel starts with several basic assumptions that ultimately lead to an equation which relates the activity coefficient of an ion in an electrolyte solution to the concentration of the other ions present. This is the relationship that, as described above, is sought to define the chemical potential change due to ion–ion interactions of the ion under scrutiny. The theoretical treatment by Debye and Hückel makes several principal assumptions:

(1) A central reference ion of a specific charge can be represented as a point charge.
(2) This central ion is surrounded by a cloud of smeared-out charge contributed by the participation of all of the other ions in solution.
(3) The electrostatic potential field in the solution can be described by an equation that combines and linearizes the Poisson and Boltzmann equations.
(4) No ion–ion interactions except the electrostatic interaction given by a 1/r^2 dependence are to be considered (i.e., dispersion forces and ion–dipole forces are to be excluded).
(5) The solvent simply provides a dielectric medium, and the ion–solvent interactions are to be ignored, so that the bulk permittivity of the solvent can be used.

In the *Debye–Hückel model*, because the central reference ion is surrounded by a countering cloud of charge, knowledge of the density of the cloud relative to the distance away from the central ion will give information about the arrangement of the ions that actually make up the cloud. Because the simplifying premise of the theory smears the individual ions into a continuum, the question of charge density in the cloud relative to the distance away from the central ion can be approached as a problem in electrostatics where the charge under consideration is a time-averaged smear. The question to be solved is How does the charge density of the cloud change with respect to the distance away from the central ion? Because the excess charge density in the smeared cloud represents a separation of charge, the relationship between the two separated charges, that is, the central ion and a specific volume element of the cloud, can be discovered by examining the electrostatic potential that exists between the two charges (Fig. 16.3).

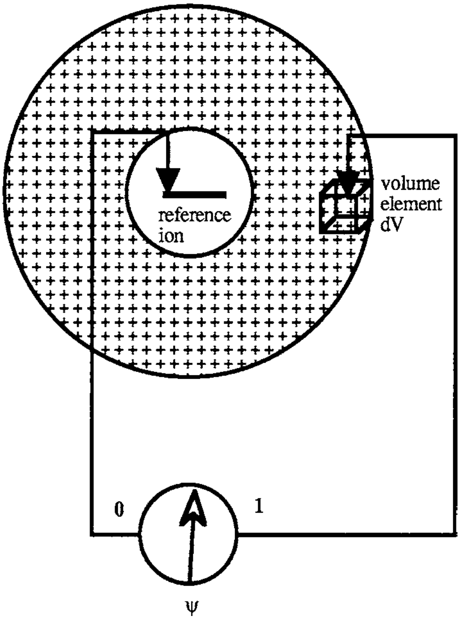


Fig. 16.3 Examining the electrostatic field existing between a reference ion and a volume element provides information about the spatial position of the charge density in the ionic atmosphere

The excess charge in a given volume, *dV*, of the cloud can be related to the electrostatic potential existing between the central ion and the small volume element under consideration. This relationship is given by *Poisson's equation*, which relates the charge distribution, ρ_r, to the electrostatic potential, ψ, in a spherically symmetrical system such as the one under discussion. The equation is written as follows:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2\frac{d\psi_r}{dr}\right) = -\frac{\rho_r}{\varepsilon_0\varepsilon} \tag{16.5}$$

This equation relates the electrostatic potential to the excess charge density in the volume element as it varies with distance from the central reference ion. The dielectric constant in this equation is usually taken as that of bulk water.

The total charge in a given volume element is found by adding together all of the charges that reside in that volume (taking into account both the number of ions and the valence number for each species). This charge can be described by

$$\rho_r = \sum n_i z_i e_{\mathrm{o}} \tag{16.6}$$

It follows therefore that the total electric charge available in a system will be described by the sum of all charged elements in solution.

Quantification of the charge in an electrolyte solution can be achieved through the use of a parameter called the *ionic strength*. The ionic strength is given by ***I***, on a molal basis where

$$I = \frac{1}{2N_A} \sum n_i^{\mathrm{o}} z_i^2 \tag{16.7}$$

and because $n_i^{\mathrm{o}} = c_i N_A$, I can be expressed in terms of concentration:

$$I = \frac{1}{2} \sum c_i z_i^2 \tag{16.8}$$

As an example, consider the ionic strength of a 1:1 electrolyte, 0.15 M NaCl solution:

$$\begin{aligned} I &= \tfrac{1}{2}\left\{\left[\mathrm{Na}^+\right](1)^2 + [\mathrm{Cl}^-]\,(-1)^2\right\} \\ &= \tfrac{1}{2}\left\{[0.15]\,(1) + [0.15]\,(1)\right\} \\ &= 0.15 \end{aligned} \tag{16.8a}$$

Taking as a second example a 1:2 electrolyte, I for a 0.15 M $\mathrm{CaCl_2}$ solution would be

$$\begin{aligned} I &= \tfrac{1}{2}\left\{\left[\mathrm{Ca}^+\right](2)^2 + [\mathrm{Cl}^-]\,(-1)^2\right\} \\ &= \tfrac{1}{2}\left\{[0.15]\,(4) + [0.15]\,(1)\right\} \\ &= 0.45 \end{aligned} \tag{16.8b}$$

It should be obvious that a relationship of ionic strength to concentration, $I = kc$, exists, where the value of k is defined by the type of electrolyte in solution, $M_x A_y$. The values of k are given in Table 16.1. As will be seen shortly, the non-ideal behavior of electrolyte solutions is related to the total number of ions (and given by the ionic strength), rather than to the actual chemical nature of the species making up the solution.

Returning now to the expression for charge excess, ρ_r, in a specific volume element, dV:

Table 16.1 Values of *k*, defining the relationship between ionic strength and concentration, depending on valence type of electrolyte

	M^{+}	M^{2+}	M^{3+}	M^{4+}
A^{-}	1	3	6	10
A^{2-}	3	4	15	12
A^{3-}	6	15	9	42
A^{4-}	10	12	42	16

$$\rho_r = \sum n_i z_i e_{\text{o}} \tag{16.6}$$

It is necessary to be able to characterize each of the elements n_i, and to do this the Boltzmann distribution law is used. The form of this law was given earlier:

$$n_i = n_i^{\text{o}} e^{-U/kT} \tag{16.9}$$

U in the case described here represents the potential energy change associated with the change in distribution of n_i particles in the volume element as distinguished from the distribution of particles in the bulk given by n_i^{o} U is a time average of all the forces that act on the particles and as such represents the sum of all forces that influence the distribution of the particles or ions. If U is positive, then the distribution of ions in the volume element is less than the bulk distribution, while a negative value for U indicates that the distribution of ions is increased relative to the bulk distribution. When U is zero, the distribution of ions in the volume element is identical to the bulk distribution. Because a central tenet of the Debye–Hückel theory is that the only interactional forces acting between the ions are electrostatic in nature and follow a $1/r^2$ relationship, the term U can be evaluated as

$$U = z_i e_{\text{o}} \psi_r \tag{16.10}$$

The Boltzmann relationship now becomes

$$n_i = n_i^{\text{o}} e^{-z_i e_{\text{o}} \psi_r / kT} \tag{16.11}$$

The distribution of the ions in the volume element as a function of the bulk distribution is now described, and Eq. (16.6) can be modified to reflect this

$$\rho_r = \sum n_i^{\text{o}} z_i e^{-z_i e_{\text{o}} \psi_r / kT} \tag{16.12}$$

This expression can be made linear if a simplifying condition is established. This is another of the assumptions on which Debye and Hückel built their model. If only cases are chosen where the electrostatic potential ψ_r is so small that the term $z_i e_{\text{o}} \psi_r / kT$ is much less than one, then the exponential in Eq. (16.12) can be rewritten as a Taylor power series:

$$e^{z_i e_{\text{o}} \psi_r / kT} = 1 - \frac{z_i e_{\text{o}} \psi_r}{kT} + \frac{1}{2} \left(\frac{z_i e_{\text{o}} \psi_r}{kT} \right)^2 \ldots \tag{16.13}$$

and all but the first two terms can be ignored. Equation (16.12) becomes

$$\rho_r = \sum n_i^{\mathrm{o}} z_i e_{\mathrm{o}} \left(1 - \frac{z_i e_{\mathrm{o}} \psi_r}{kT}\right) \tag{16.14}$$

or

$$\rho_r = \sum n_i^{\mathrm{o}} z_i e_{\mathrm{o}} - \sum n_i^{\mathrm{o}} \frac{z_i^2 e_i^2 \psi_r}{kT} \tag{16.15}$$

The first term is the same as Eq. (16.4), which gave the total charge of the solution, and by electroneutrality must be equal to zero. Equation (16.15) therefore simplifies to

$$\rho_r = -\sum \frac{n_i^{\mathrm{o}} z_i^2 e_i^2 \psi_r}{kT} \tag{16.16}$$

This is the *linearized Boltzmann equation.*

By combining the linearized Boltzmann equation and the Poisson equation, each relating the charge density, ρ_r, in the volume element, dV, to the distance, r, from the reference ion, the *linearized Poisson–Boltzmann equation* is obtained:

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d\psi_r}{dr}\right) = \frac{1}{\varepsilon_{\mathrm{o}} \varepsilon kT} \sum n_i^{\mathrm{o}} z_i^2 e_i^2 \psi_r \tag{16.17}$$

If all of the right-hand terms are collected into a single variable, κ^2, the expression may be rewritten as

$$\frac{1}{r^2}\frac{d}{dr}\left(r^2 \frac{d\psi_r}{dr}\right) = \kappa^2 \psi_r \tag{16.18}$$

The variable κ is important and will be discussed shortly.

Equation (16.18) can be solved by considering the boundary conditions that derive from the consideration of the ion as a material point charge with a field that extends to infinity. Consequently, the problem is integrated from $r = 0$ to $r = \infty$ and the result is the following:

$$\psi_r = \frac{z_i e_{\mathrm{o}}}{4\pi \varepsilon_{\mathrm{o}} \varepsilon r} e^{-\kappa r} \tag{16.19}$$

This equation can be approximated and expanded to describe a contribution to the electrostatic field by the central ion and the cloud that surrounds it. The ion's contribution is $z_i e_{\mathrm{o}}/4\pi\varepsilon_{\mathrm{o}}\varepsilon r$ while the contribution of the cloud is $z_i e_{\mathrm{o}}/4\pi\varepsilon_{\mathrm{o}}\varepsilon r L_D$, where L_{D} represents a distance from the central ion and is equal to κ^{-1} as described in the following paragraph. The field contributed by the cloud of charge must counter that contributed by the ion, and the expansion of Eq. (16.19) confirms this

$$\psi_r = \frac{z_i e_{\mathrm{o}}}{4\pi\varepsilon_{\mathrm{o}}\varepsilon r} - \frac{z_i e_{\mathrm{o}}}{4\pi\varepsilon_{\mathrm{o}}\varepsilon L_D} \tag{16.20}$$

The total field given by ψ_r then is comprised of contributions from the ion and from the charge cloud:

$$\psi_r = \psi_{\mathrm{ion}} + \psi_{\mathrm{cloud}} \tag{16.21}$$

It follows that

$$\psi_{\mathrm{cloud}} = -\frac{z_i e_{\mathrm{o}}}{4\pi\varepsilon_{\mathrm{o}}\varepsilon L_D} = -\frac{z_i e_{\mathrm{o}}}{4\pi\varepsilon_{\mathrm{o}}\varepsilon\kappa^{-1}} \tag{16.22}$$

A brief survey of the problem so far will be valuable. In an effort to determine the alteration in the chemical potential of an ionic species in solution due to the interactions between ions, a model first proposed by Debye and Hückel has been presented. This model is based on the assumption that the most significant interactional forces can be described in terms of electrostatics. A central reference ion is chosen with a charge $z_i e_0$ and is considered as a point charge. Since it is a point charge, it generates a potential field that extends into space in a spherically symmetrical manner. The remaining ions in the solution constitute an electrical charge that surrounds the central ion and will be arranged so that another electrostatic field is generated. This field will act to neutralize the potential field created by the central ion. In the Debye–Hückel model, the surrounding ionic charge is treated as a continuum of charge much like that described previously in the discussion of the Born model. By treating the question of alteration in chemical potential or activity in this fashion, the problem becomes one in electrostatics. If the interactional forces that are described by the electrostatic potential fields between the reference ion and the surrounding charge cloud can be defined, then the change in chemical potential can subsequently be found. The contributions to the electrostatic field can be found for the central ion since the field is determined simply by the charge of the ion, $z_i e_0$. In the case of the field contributed by the countering charge cloud, the field could be calculated after the charge density in a specific volume element under study, *dV*, was found. Through the use of the Boltzmann equation, the distribution of charge in the volume element was defined, and then a linear relationship was derived by making an assumption that thermal forces were much more significant than electrostatic forces in the distribution of ions. If this were not the case, the ions making up the cloud would condense around the central ion and form a crystal. At this point, Eq. (16.20) relates the contributions of the central ion and the surrounding cloud of charge to the electrostatic potential, ψ_r, at a particular distance from the central ion. A very important feature of the derivation to this point is the result that the field contributed by the charge cloud is related to the parameter, L_{D}, which has units of length. The physical interpretation of this mathematical expression is that the ionic atmosphere can be replaced by a charge at the distance L_{D} from the central ion. L_{D} is equal to κ^{-1} and is often called the *Debye length* or the *effective radius* of the charge atmosphere surrounding the central ion.

The effective radius is obviously an important physical aspect of an ion in relationship to other ions in solution. Earlier in the derivation, the term κ was created by grouping a series of constants together. What are the parameters of κ that have an effect on the thickness of the ionic atmosphere? Referring to Eq. (16.17), κ^2 can be written as follows:

$$\kappa^2 = \frac{1}{\varepsilon_0 \varepsilon kT} \sum_i n_i^{\mathrm{o}} z_i^2 e_i^2 \tag{16.23}$$

Following rearrangement and writing κ^2 on a molar basis:

$$\kappa^2 = \frac{e_{\mathrm{o}}^2 N_A}{\varepsilon_0 \varepsilon kT} \sum_i n_i^{\mathrm{o}} z_i^2 \tag{16.24}$$

The last term in Eq. (16.24) should be familiar as closely related to the definition of ionic strength:

$$I = \frac{1}{2} \sum c_i z_i^2 \tag{16.8}$$

Therefore, by substitution, κ^2 can be expressed in terms of ionic strength:

$$\kappa^2 = \frac{2 e_{\mathrm{o}}^2 N_A}{\varepsilon_0 \varepsilon kT} I \tag{16.25}$$

k^{-1} or L_D can be written as follows:

$$\kappa^{-1} = L_D = \left(\frac{\varepsilon \varepsilon_0 kT}{2 e_{\mathrm{o}}^2 N_A I} \right)^{1/2} \tag{16.26}$$

This means that the effective radius of the ionic atmosphere is inversely related to the ionic strength of the solution. The radii of ionic atmospheres for electrolytes of

Table 16.2 Radii in nm of ionic atmosphere ($L_\mathrm{D} = \kappa_{-1}$) for electrolytes of different valence types at various concentrations at 298 K

	Valence type				
Molality	1:1	1:2	1:3	2:2	2:3
10^{-5}	96.1	55.5	39.2	48.00	24.80
10^{-4}	30.40	17.50	12.40	15.20	7.80
10^{-3}	9.61	5.55	3.92	4.80	2.48
10^{-2}	3.04	1.75	1.24	1.52	0.78
10^{-1}	0.96	0.56	0.39	0.48	0.25
1	0.30	0.18	0.12	0.15	0.08

several valence types at several concentrations are given in Table 16.2. The physical ramifications of this relationship will be examined shortly.

The link between the ionic atmosphere and ψ_r and the effect on the chemical potential of the ion as expressed by the activity coefficient is developed by employing a charging process similar to that described in the formulation of the Born model. To find the potential energy of the ion–ion interaction, the ion is first considered to be discharged and then is brought to its charge, $z_i e_o$. The work associated with charging the ion is the potential energy and gives the alteration in chemical potential for the system. Earlier, this relationship between the change in chemical potential and the charging process was given as

$$\Delta\mu_{i-I} = N_A \frac{Z_i e_o}{2} \psi \tag{16.3}$$

ψ is now well described (Eq. (16.11)), and this equation can be rewritten as

$$\Delta\mu_{i-I} = -\frac{N_A \left(z_i e_o\right)^2}{8\pi\varepsilon_o \varepsilon \kappa^{-1}} \tag{16.27}$$

Previously, the activity coefficient was introduced and defined in terms of a difference between the ideal and real chemical potentials:

$$\Delta U_{(\text{real - ideal})} = (\mu_o + RT \ln X_i + RT \ln \gamma_i) - (\mu_o + RT \ln X_i) = RT \ln \gamma_i \tag{16.28}$$

An ideal system is defined as being composed of non-interacting particles, while for a real system the interactions between particles must be taken into account. In the Debye–Hückel model, the interactions are the ion–ion interactions, and consequently $\Delta U_{(\text{real}-\text{ideal})}$ is $\Delta U_{i\text{–}I}$. Combining Eq. (16.27) with Eq. (16.28) therefore gives the result

$$-\frac{N_A \left(z_i e_o\right)^2}{8\pi\varepsilon_o \varepsilon \kappa^{-1}} = RT \ln \gamma_i \tag{16.29}$$

At this stage, a theoretical relationship has been derived, based on a particular model, that will predict the activity coefficient for a single ionic species in solution. Originally, the concept of the activity coefficient was introduced as an empirical device to allow continued use of the derived ideal equations for thermodynamic systems, and so it is grounded in experiment. The benefit in deriving Eq. (16.29) lies in the fact that, if the theory accurately predicts the activity coefficients found experimentally, it is validation of the structural understanding implicit in the theoretical model. Therefore, before going further, it is crucial to know just how well this model and experiment agree.

16.3 The Predictions Generated by the Debye–Hückel Model Can Be Experimentally Evaluated

As is so often the case, the simplistic elegance of a model is not so easily tested in the laboratory. A problem similar to the one that described in the experimental validation of the Born model exists here. To measure the interactions, the solution must remain uncharged and electroneutral. If a single ionic species is added, this condition cannot be met, and the measurement will include the confounding interaction of the reference ion with an electrified solution. For this reason, activity coefficients cannot be experimentally measured for single ions, but instead the activity coefficient of a net electrically neutral electrolyte is used. This consideration leads to the *mean ionic activity coefficient*. If a mole of electrolyte MA is added into solution, the chemical potential for the new solution is described by the sum of each ion's chemical potential, i.e.

$$\mu_{M^+} + \mu_{A^-} = \left(\mu^{\mathrm{o}}_{M^+} + \mu^{\mathrm{o}}_{A^=}\right) + RT\ln\left(X_{M^+}X_{A^-}\right) + RT\ln\left(\gamma_{M^+}\gamma_{A^-}\right) \qquad (16.30)$$

However, this equation actually describes a 2 M solution, since one mole of each ion has been added. In order to determine the free energy contribution from one mole of both ions comprising the electrolyte, Eq. (16.30) must be divided by two:

$$\frac{\mu_{M^+} + \mu_{A^-}}{2} = \frac{\left(\mu^{\mathrm{o}}_{M^+} + \mu^{\mathrm{o}}_{A^=}\right)}{2} + RT\ln\left(X_{M^+}X_{A^-}\right)^{1/2} + RT\ln\left(\gamma_{M^+}\gamma_{A^-}\right)^{1/2} \qquad (16.31)$$

This equation defines a mean quantity derived from the average contributions of each species to the system as a whole. Each of these average quantities is measurable. These average quantities are called *the mean chemical potential*, *bf* $\mu_{\pm}$, *the mean standard chemical potential*, $\mu_{0\pm}$, *the mean mole fraction*, $\boldsymbol{X}_{\pm}$, and the *mean ionic activity coefficient*, $\gamma_{\pm}$. The general formula relating the measurable mean ionic activity coefficient to the activity coefficient for each individual ion is

$$\gamma_{\pm} = \left(\gamma^{v^+}\gamma^{v^-}\right)^{1/v} \qquad (16.32)$$

where

$$v = v^+ + v^- \qquad (16.33)$$

Using this relationship and the Debye–Hückel formulations for γ^+ and γ^-, an equation is found that relates the theoretical calculation of the activity coefficient to the experimentally available mean ionic activity coefficient:

$$\log\gamma_{\pm} = -A\left(z_+ z_-\right) I^{1/2} \qquad (16.34)$$

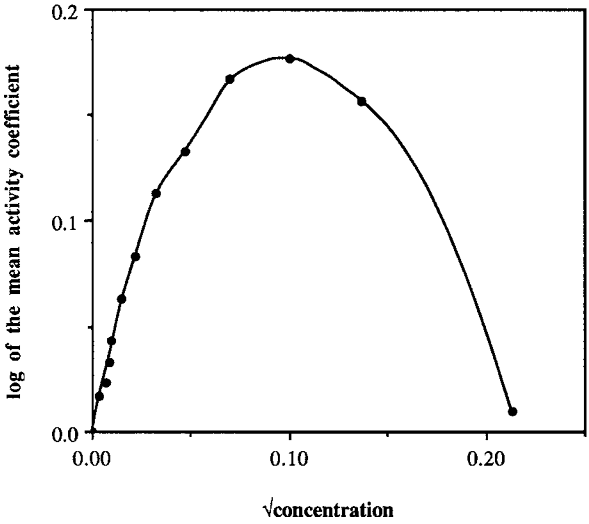
Table 16.3 Values of *A* for water at various temperatures

Temperature (K)	A (dm$^{3/2}$ mol$^{-1/2}$)
273	0.4918
293	0.5070
298	0.5115
303	0.5161
313	0.5262
323	0.5373

For water, the constant *A* has values as shown in Table 16.3 with units of dm$^{3/2}$ mol$^{-1/2}$. In SI units, the value of *A* for water at 298 K is 1.6104 × 10^{-2} m$^{3/2}$ mol$^{-1/2}$. Equation (16.34) is called the *Debye–Hückel limiting law* and is the link between experimental values and the model developed so far.

When the comparison between theory and experiment is made (Fig. 16.4), the limiting law is found to be quite accurate in solutions of 1:1 electrolytes of concentrations no greater than 0.01 N. However, for more concentrated solutions the theoretical and experimental values diverge significantly and even in moderately concentrated solutions of only 1 N, the mean activity coefficient will begin to rise. Additional problems exist; variations in the mean activity coefficient are found with electrolytes of different valence types (i.e., 1:1, 1:2 electrolytes), as well as with different electrolytes of the same valence type.

Fig. 16.4 Plot of experimental values for the mean activity coefficient for NaCl at various concentrations



The limits of the Debye–Hückel limiting law are significant for the biological worker. The ionic strengths of biological fluids are considerably higher than the useful limit of Eq. (16.34). It is clear that the ionic strength of solutions in vitro plays a significant role in cellular and biochemical behavior. It is relevant therefore to attempt to gain a greater insight into the behavior of more concentrated solutions. The first step will be to examine the assumptions on which the limiting law is based and to see if modifications of these assumptions will lead to greater agreement between theory and experiment.

16.4 More Rigorous Treatment of Assumptions Leads to an Improved Performance of the Debye–Hückel Model

One of the basic tenets of the Debye–Hückel model was the treatment of the central ion as a point charge. Obviously, this is a drastic oversimplification, since ions are finite in size and must occupy a certain space. Could this assumption account for some of the problems with the limiting law? Consider that the limiting law linearly relates the mean ionic activity coefficient to the square root of the ionic strength. Earlier (Eq. (16.26)), it was shown that the radius of the ionic atmosphere is inversely related to the concentration through the constant κ. The ionic atmosphere is relatively large for a dilute solution and becomes progressively more compact as the concentration increases. Table 16.2 shows that the radius of the ionic atmosphere for a 1:1 electrolyte such as NaCl is 96 nm at a concentration of 10^{-5} N, yet at a concentration of 10^{-1} N, it has narrowed 100-fold to 0.96 nm. Since an ion is approximately 0.2 nm in diameter, the cloud at a concentration of 1 N will be of a similar dimension to the ion. This is a problem since the point charge approximation depends on the idea that the radius of the ionic atmosphere is much greater than the ionic radius, a simplification that becomes increasingly inaccurate as the dimension of the ionic cloud diminishes. The Poisson formulation also is based on the premise that discrete ions can be treated as a smeared charge, an assumption that becomes unsustainable when the ions are concentrated into a much smaller space. At high concentration, the ions respond to other ions as discrete charges, and the Poisson equation becomes less valid.

The point charge assumption was made when the limits were chosen in solving Eq. (16.19), and it is here that the finite size of the ion will be introduced. The lower limit of integration will no longer be zero, since the central ion occupies this point. The lower level is more logically chosen to start at the edge of the finite-sized ion, a specific distance, a, from the center point of the ion. The upper limit can still be considered to be at infinity, and the result of the new equation is

$$\psi_r = \frac{z_i e_o}{4\pi\varepsilon_o\varepsilon}\frac{e^{\kappa a}}{(1+\kappa a)}\frac{e^{-\kappa r}}{r} \tag{16.35}$$

This defines the potential ψ_r at the distance r from a central ion of finite size a. A procedure similar to that used to obtain Eq. (16.34) may now be applied to yield the following formula that relates the mean ionic activity coefficient to the finite-central-ion model:

$$\gamma_{\pm} = \frac{-A\,(z_{+}z_{-})\,I^{1/2}}{1 + BaI^{1/2}} \tag{16.36}$$

The constant A is the same as that in the limiting law (Eq. 16.34), and B can be written as Eq. (16.37) and has the value of 3.291×10^{9} m^{-1} $\mathrm{mol}^{-1/2}$ $\mathrm{kg}^{1/2}$ for water at 298 K:

$$B = \left(\frac{2e_0^2 N_A}{\varepsilon_0 \varepsilon kT}\right)^{1/2} \tag{16.37}$$

As the solution becomes increasingly dilute and $I \rightarrow 0$, the second term in the denominator of Eq. (16.36) approaches zero, and Eq. (16.36) reduces to Eq. (16.34). The physical interpretation of this result is that at very dilute concentrations the radius of the ionic atmosphere is so much larger than the finite radius, a, of the central ion that the ion can be effectively treated as a point charge. However, as the concentration increases, the finite size of the ion must be taken into account.

In fact, the value of a, the effective radius of the central ion, is difficult to know exactly but is usually considered to vary between 0.3 and 0.5 nm. These values are found from experiment and represent distances of closest approach between two ions. When values in this range are used, Eq. (16.36) shows a much greater accuracy in predicting the mean activity coefficient up to a concentration of 0.1 N. Furthermore, because the model of the ion of finite size takes into account the actual identity of the ions, at least as reflected in their effective radii and not just in their valence types, this model can account for the differences in observed activity coefficients for electrolytes of the same valence type such as NaCl and KCl.

The Debye–Hückel theory as quantified in Eq. (16.36) is truly a remarkable success, and the model it describes is an important one both conceptually and practically. However, the disparity between the theoretical predictions and the experimental values for mean activity coefficients becomes a serious one in the range of ionic strengths of many biologically relevant solutions. Are there further modifications that can be considered to explain the experimental values? The logical next step is to examine the assumptions on which the model was originally constructed. Substitution of increased detail for the simplifications, as in the case of the addition of finite ion size, will lead to a more sophisticated model. In the case of this model, the original assumptions were

(1) Point charge representation of the central ion.
(2) Availability of all ions in solution to participate freely with the central reference ion.
(3) Valid use of the linearized Poisson–Boltzmann equation.
(4) Only ion–ion interactions with a $1/r^2$ dependence considered.
(5) No ion–solvent interactions.

So far, it has been shown that the first and third assumptions may be oversimplifications in systems of practical biological relevance. The addition of the finite radius, *a*, improved the agreement between theory and experiment. Will consideration of the other approximations provide greater understanding of the shape of the curve found experimentally as shown in Fig. 16.4?

16.5 Consideration of Other Interactions Is Necessary to Account for the Limits of the Debye–Hückel Model

In Chapter 15, the interactions between ions and their aqueous solvent were considered in some detail. These interactions have been completely ignored up to this point. We now ask: What effect might the interaction between the ions and water have on the mean activity coefficient? One of the most important aspects of the ion–solvent interactions was the formation of hydration sheaths. Molecules of water that are tightly bound to an ion in the hydration sheath will no longer be free to interact with new electrolyte added. By an effect analogous to the salting-out effects discussed earlier, as an increasing concentration of electrolyte is added, a decrease in the effective concentration of the water occurs. For a 1 M solution of LiCl, whose primary hydration number can be considered as 7, the number of moles of water immobilized by the addition of the electrolyte will be 7. This represents a 12.6% drop in the effective solvent concentration. A 3 M LiCl solution, commonly used in the extraction of RNA from cells, effectively removes 21 mol of solvent or 38% of the water available. As the effective concentration of the solvent falls, the effective concentration of the electrolyte increases. Since it is the activity coefficient that relates actual concentration to the effective concentration, the activity coefficient of the water will decrease, and the activity coefficient of the LiCl will increase. This effect accounts partly for the increase seen in the mean activity coefficient at increasing concentrations.

16.5.1 Bjerrum Suggested That Ion Pairing Could Affect the Calculation of Ion–Ion Interactions

The model as it now stands has been modified to take into consideration that (1) the ionic radii are not point charges, and (2) there are significant ion–solvent interactions. Another basic assumption is that all of the ions in solution are free to contribute to the ionic atmosphere. The motions and positions of the ions are considered random, secondary to thermal forces, except as restricted by the coulombic forces that act between the central ion and the localized ionic atmosphere. Even though it is assumed that the thermal forces are generally much greater than the coulombic forces (Eq. (16.13)), there is a probability that an ion acting as a member of the ionic atmosphere may get close enough to the central ion that it will become locked in a coulombic embrace, and no longer then will either ion be independent of the other. When such an event occurs, Bjerrum suggested that an *ion pair* would be formed. When an ion pair forms, the number of free ions in solution decreases. The electric charge of an ion pair is less than that of a free ion; in fact, the charge of an ion pair is frequently zero. Any property of an electrolytic solution that depends on the number of free charged particles (i.e., the ionic strength) will be affected by ion pairing. The probability that ion pairing will occur depends on whether the electrostatic fields of two mutually attractive ions will overcome thermal randomizing forces. A minimum distance, *q*, may be defined at which the mutual attraction of the oppositely charged ions will be comparable to their thermal energy:

$$q = \frac{|z_+|\,|z_-|\,e_o^2}{8\pi\varepsilon kT} \tag{16.38}$$

For ion pair formation to occur, the physical size of the ions must be taken into account. Obviously, if the ions are too large to approach one another within the distance *q*, ion pairing will be impossible. If the sum of the effective radii of the ions, *a*, is less then *q*, then ion pairing can occur, i.e.

If $a < q$ then ion pairing occurs,
if $a > q$ then no ion pairing.

For electrolytes in water at 298 K, *q* can be written as

$$q = 0.336\,|z_+|\,|z_-|\,\text{nm} \tag{16.39}$$

For 1:1 electrolytes in water, virtually no ion pairing occurs, since *a* is usually between 0.3 and 0.5 nm and these electrolytes are completely dissociated. However, ions of higher valence, such as Ca^{2+} and Mg^{2+}, may form ion pairs in aqueous solution.

We have treated the subject of ion–ion interaction at a somewhat elementary level in order to provide a flavor of the modeling process and to give an adequate foundation for further study. Biological systems often need a more extensive treatment especially with respect to both high concentrations and the polyelectrolyte nature of macromolecules. Several of the articles in the further reading list are a good place to start.

Further Reading

Anderson C.F. and Record M.T. Jr. (1995) Salt-nucleic acid interactions, *Annu. Rev. Phys. Chem.*, **46**:657–700. (A sophisticated but readable treatment of solute-solute interactions in solutions of nucleic acid polyelectrolytes.)

Bockris J.O’M, Reddy A.K.N., and Gamboa-Aldeco M. (1998) *Modern Electrochemistry*, 2nd edition. Kluwer Academic/Plenum, New York.

Koryta J. and Dvorak J. (1993) *Principles of Electrochemistry*, 2nd edition. Wiley, New York.

Wolynes P.G. (1980) Dynamics of electrolyte solutions, *Annu. Rev. Phys. Chem.*, **31**:345–376. (A good introduction to the modern questions of electrolyte theory that rest on the earlier work of Debye and Huckel.)

Problem Sets

1. What is the ionic strength of the serum component of blood? What is the osmolarity of this component?
2. Calculate the effective radius of the ionic atmosphere for ions in the serum. Use the ions listed in question and base your answer on the work derived in question 65.
3. Name two situations in biological systems where ion pairing as described by Bjerrum may be of consequence. Justify your answer.
4. What is the ionic strength of the following?

   (a) 0.5 M KCl
   (b) 0.5 M $CaCl_2$
   (c) 0.1 M K_2MNO_4
   (d) 0.3 M H_2CO_3
   (e) 0.01 M $Ba(OH)_2$

5. Lithium chloride and urea are used in the isolation of RNA. Based on your knowledge of aqueous solutions, why have these substances been found so effective?
6. What is the radius of the Debye atmosphere for the cations in question 4 at 298 K? at 310 K? at 400 K? at 274 K?
7. A new potassium channel is discovered whose conductivity is proportional to the activity of extracellular K^+. Using the Debye–Hückel limiting law, draw a graph relating the channel conductivity to the concentration of serum potassium (2.5–6 meq/l).
8. The relationship between the channel conductivity and potassium concentration is found to be very sensitive to serum calcium. Why?

Chapter 17
Lipids in Aqueous Solution

Contents

17.1 Biological Membranes Form at the Interface Between Aqueous and Lipid Phases

Thus far, we have explored the properties of inorganic aqueous solutions and of bulk water. Yet cells and biological systems are not comprised of simple homogeneous solutions or bulk aqueous phases. We have emphasized that a fundamental conceptual and structural aspect of biological organization is the cell membrane. Cells are defined physically and functionally in relationship to their environment by the cell membrane. The organization of biological systems, and certainly eukaryotic cells, depends on compartmentalization of cellular functions. This compartmentalization, which allows varied microenvironments to exist in close proximity, is dependent on the membranes that constitute the boundaries of the compartmentalized organelles. Virtually all biological membranes are comprised of lipid molecules arranged in bilayers, with inserted or attached protein and carbohydrate molecules

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_17,

playing a variety of roles at or in the lipid phases. While membranes can be formed from a wide variety of polymeric substances including proteins and carbohydrates, biological membranes are generally comprised of lipid molecules which form the membrane structure itself. Although membranes in cells are comprised primarily of hydrophobic lipid elements, they almost always separate two phases whose predominant species is water. Generally, these aqueous-dominated phases are treated as aqueous solutions. Biological membranes are characterized as being permeable to some but not to all of the components of these solutions. The arrangement of these aqueous–lipid–aqueous "phases" leads to a generalized mechanism through which the cell can perform a wide variety of tasks that allow it to sense, judge, and respond to its environment. In other words, the membrane not only defines where the cell as "self" and the environment as "not self" begins and ends, but also it allows the cell to collect information and energy from the environment, process these inputs, and respond in a variety of output patterns. The membrane either is active in or is the site of action for physical separation and protection; mediation of the flow of chemical species into and out of the cell; accumulation and conversion of energy and electric charge for the cell; and information processing and exchange between cellular elements and the environment, both local and remote. How can we understand membrane formation? We start by exploring the interactions of water with nonpolar molecules and then focus on certain biophysical and biochemical aspects of lipid membrane formation from which a more complete description of the cell can eventually emerge.

17.2 Aqueous Solutions Can Be Formed with Small Nonpolar Molecules

The hydration of small nonpolar molecules is qualitatively different from the electrostatically dominated interaction described for polar molecules. In nonpolar solvation, there is "hydrophobic hydration" or structuring of the solvent around the solute molecule. This is an entropically unfavorable step, and therefore $\Delta S_{\text{solvation}}$ for nonpolar molecules is always negative. When an increasing concentration of nonpolar molecules is introduced into the aqueous system, there is a strong gain in entropy because as nonpolar molecules associate with one another there is a release of the water molecules previously involved in the entropically unfavorable hydration of these molecules. As discussed previously this is the *hydrophobic effect*.

When a nonpolar solute is small, that is, similar in size to H_2O, the thermodynamic properties (Table 17.1) are remarkably independent of the other properties of the solute. Compared with the solute-nature-dependent heat of solvation of ions just discussed, this is a major difference. The heats of solvation of small nonpolar molecules reflect an interaction energy common to all of these solutes. This energy is associated with the insertion of these small compounds into the partially ordered structures of liquid H_2O, which become more stable as a result of the insertions.

Initially, it was thought that each of the entities listed in Table 17.1 would fit into the space within the center of the hexagonal arrays (cf. Chapter 14), much as molecules of H_2O fit into the space in liquid water. The diameter of these

Table 17.1 Enthalpies of hydration for selected small nonpolar molecules

Compound	$\Delta H_{\text{solvation}}$ (kJ/mol)
Ar	−69.5
Kr	−58.2
H_2S	−69.0
PH_3	−62.8
SO_2	−69.5
CH_4	−60.7
C_2H_2	−62.8
C_2H_4	−62.8
C_2H_6	−62.8
CH_3Cl	−62.8
CH_2SH	−69.5

inserted nonpolar substances would then have to be under approximately 0.54 nm. However, evidence points to a stabilized H_2O structure around the nonpolar solute with more pentagonal rather than hexagonal order. Each of the water molecules in the pentagonal array found in the vicinity of nonpolar molecules is still tetrahedral, that is, still has a coordination number of 4, but the orientation of the individual H_2O molecules is different and the tetrahedral angle is deformed, with the hydrogen bonds no longer being linear (Fig. 17.1). Thus the hydrogen bonds in these disturbed arrays are slightly bent and strained.

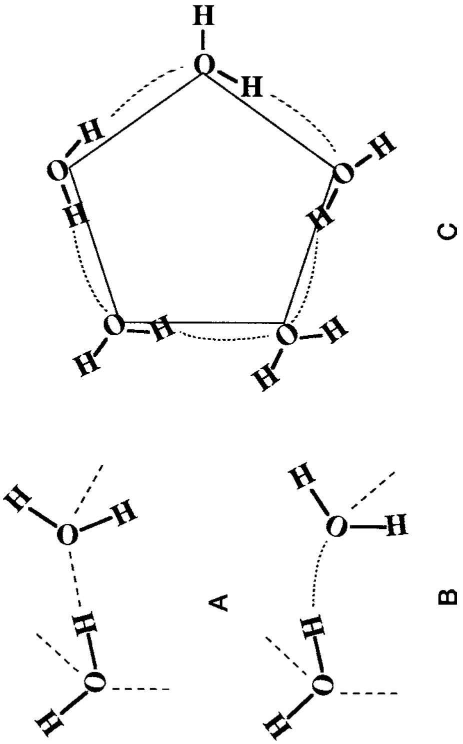


Fig. 17.1 Comparison of hydrogen-bonded water arrays: (**a**) hydrogen bonding in bulk water; (**b**) water molecules in a nonpolar array with deformation of the previously linear hydrogen bonds; (**c**) the resultant pentagonal structure in the presence of a nonpolar molecule

The pentagonal structures have been called “icebergs” by Pople and are of definite size and spacing. Their stabilization requires the inserted nonpolar molecules. Removal of the solute or addition of an ionic species which effectively disrupts the solvent structure causes reversal of the solubilization and precipitation of the solute. The procedure whereby ionic species are added to a solution containing a solvated nonpolar species in order to remove it from solution is called “salting out.” Such a technique is frequently used in biochemical separation techniques where the solute is a macromolecule.

The stabilization of H_2O structure by a nonpolar solute leads to a higher heat capacity and higher melting temperatures for these solutions when compared to bulk water. Furthermore, the pentagonal structures shown in Fig. 17.2 have very defined spaces within the pentahedral array. Thus, a nonpolar residue which has been “solubilized” will have a predictable number of H_2O molecules associated with it, giving it the regularity of a crystal and the stoichiometry of a hydrate. The smallest hydrate has eight possible insertion sites for small nonpolar solutes (two sites have 0.52 nm diameters and six sites are 0.59 nm in diameter). This array can include 46 molecules of H_2O. If the solute is larger and cannot fit into this array, there are a whole series of regular structures called *clathrates* with a defined number of solute and solvent sites. The latter also have a regular pentagonal array, which corresponds to a stable hydrate in which there is a given stoichiometry between specific nonpolar solute molecules and H_2O molecules. The actual existence of such clathrates, as distinguished from a nonstoichiometric highly ordered structure around a large nonpolar solute, is still the

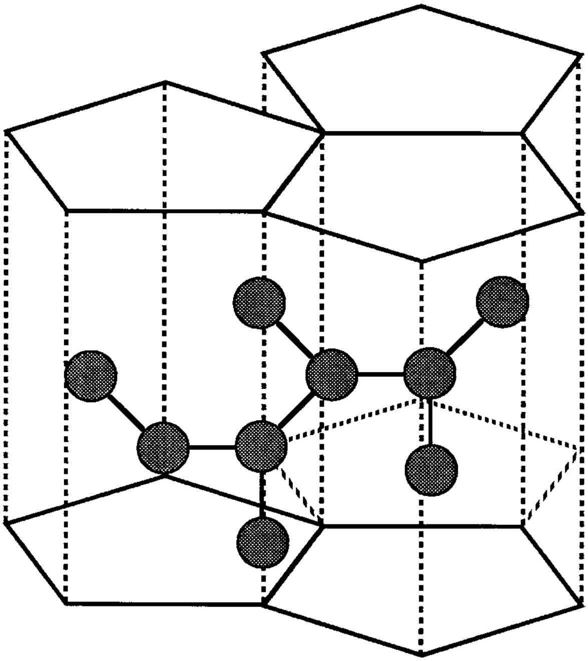


Fig. 17.2 Structure of a pentagonally organized water cavity surrounding a large solute cavity

subject of some debate and is discussed in greater detail in the references at the end of the chapter. For our purposes it will be adequate to appreciate that in the presence of nonpolar solutes, water forms a somewhat different and inherently more strained regular structure in which the solute is accommodated. The solute molecules are in close proximity to the solvent, approaching their van der Waals radii, but the solute develops no enthalpic interactions with the water molecules. As will be seen shortly, H_2O plays a major role in changes in conformation of such molecules.

17.3 Aqueous Solutions of Organic Ions Are an Amalgam of Ion-Solvent and Nonpolar Solute Interaction

The effects of the disruption of the normal hydrogen-bonded structures in bulk water by small inorganic ions have been described. The situation is different for nonpolar organic ions. These include the small aliphatic or aromatic quaternary ammonium salts and the large detergents or charged lipids. Since small and large nonpolar ions behave differently, they will be considered separately.

17.3.1 Solutions of Small Organic Ions

Small organic ions carry their single charge on an otherwise hydrophobic moiety. With respect to thermodynamic parameters, organic ions are distinct from inorganic ions. These distinctions result from the fact that the disruption of hydrogen bonding in the aqueous solvent occurs differently. The association between an inorganic ion and the irrotational water layer is a stronger interaction than an ordinary hydrogen bond, but a small organic ion is essentially incapable of such bonding. A small organic ion tends to organize its aqueous sheath as if it were an uncharged nonpolar solute. Therefore, the hydration sheath around the exterior of such an organic ion will tend to have a strongly hydrogen-bonded pentagonal structure. The positive concentration dependence of the heat capacity of solutions of quaternary ammonium salts (Fig. 17.3), for example, indicates that H_2O is more ordered around these tetraalkylammonium ions than around inorganic ions (compare to Fig. 15.16). The complex nature of the heat capacity curve for the solutions of organic ions has been difficult to explain; it implies that at low concentrations of solute the nature of the interactions is mainly hydrophobic hydration. At higher concentrations, these solutions begin to resemble inorganic ionic ones, and C_p, having gone through a maximum, begins to decrease. It is not yet possible to predict the values of ΔH and ΔS of solvation for these nonpolar ions, although it is clear that C_p for nonpolar ions is highly dependent on temperature, as well as solute concentration. In the region in which the hydrophobic solvation of the solute predominates, the increased order imposed upon the molecules of the aqueous sheath means that more, not less, energy is required to heat the sheath water.

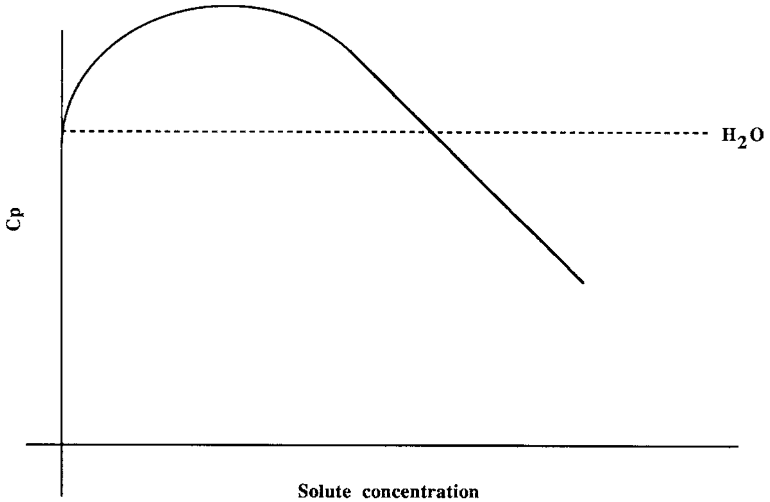


Fig. 17.3 Plot of the heat capacity (C_p) for a solution of an organic ion in water as a function of concentration

While the aliphatic or aromatic group-substituted quaternary ammonium ions behave more nearly like nonpolar solutes with respect to the structure of the aqueous hydration sheath, solutions of these ions will exhibit conductance properties dependent on the hydrodynamic properties and the presence of a net charge on the solvated entity, as shown in Table 17.2.

Table 17.2 Limiting ionic conductivity in aqueous solutions at 298 K

Ion	Conductivity (Ω^{-1} m^2 × 10^{-4})	Ion	Conductivity (Ω^{-1} m^2 × 10^{-4})
Li^+	38.69	$(CH_3)_4N^+$	44.42
Na^+	50.11	$(C_2H_5)_4N^+$	32.22
K^+	73.50	$(C_3H_7)_4N^+$	23.22
Rb^+	77.20	$(C_4H_9)_4N^+$	19.31
Cs^+	77.29		

17.3.2 Solutions of Large Organic Ions

Large organic ions that contain a single-charged entity at one end of a long aliphatic chain are amphiphilic molecules and are called soaps or detergents. Some lipids fall into this category. These molecules are distinct because their aliphatic portions will not interact with H_2O yet tend to form hydrophobic interactions with one another. Due to entropic driving forces, the aliphatic portions exclude H_2O and form van der Waals bonds, while the charged ends form hydrogen bonds and electrostatic interactions with the aqueous milieu. The effect of these very different forces on the structure of the solute, as well as its interaction with solvent, is strongly concentration dependent.

Except in very dilute solutions, these forces will cause amphiphilic molecules to associate with each other into structures called *micelles* (Fig. 17.4). In micelles, the hydrophobic portions of the molecules are separated from the aqueous solvent by a "self-generating" hydrophobic milieu, while the polar groups face the solvent and interact with the water molecules. If the aliphatic chains are short, the hydrocarbon region is not structured at room temperature, and the interior of the micelle, the hydrocarbon region, can be a pure liquid phase. No hole can exist at the center of a micelle, which is a hydrophobic region. Therefore, hydrophilic entities cannot be accommodated, but a small hydrophobic entity could be included and carried within the micelle core. The predicted size and shape of micelles depend strongly on the length and degree of unsaturation of the alkyl chains. For longer alkyl chains and for greater degrees of asymmetry, the resulting micelle may no longer be spherical but can instead be ellipsoidal or resemble a flat disk. When the disk becomes large enough and flat enough, the amphiphiles essentially form a bilayer (vide infra).

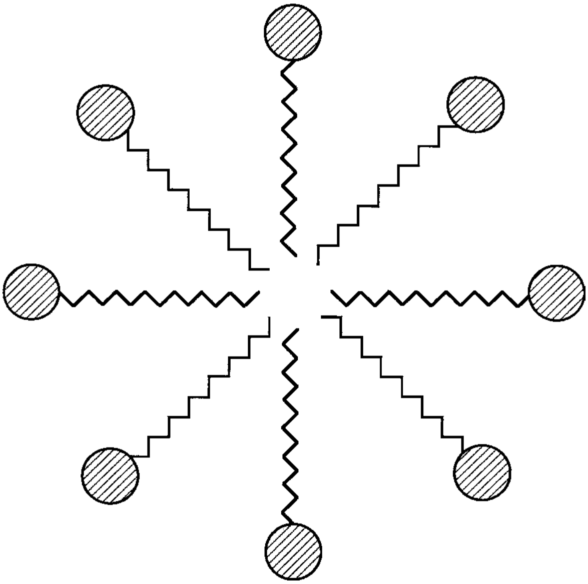


Fig. 17.4 Arrangement of large organic ions into micelles. The polar head groups form the outer hydrophilic surface that protects the hydrophobic aliphatic chains from having to interact with water

The concentration at which the change from a solubilized monomeric form to a micelle occurs is called the *critical micelle concentration*. It tends to be a convenient point of reference for the formation of the micellar phase. The actual critical micelle concentration depends strongly on the nature and length of the alkyl portion. The critical micelle concentration also depends on the charge and the nature of the polar end. The critical micelle concentration is usually determined graphically for a given

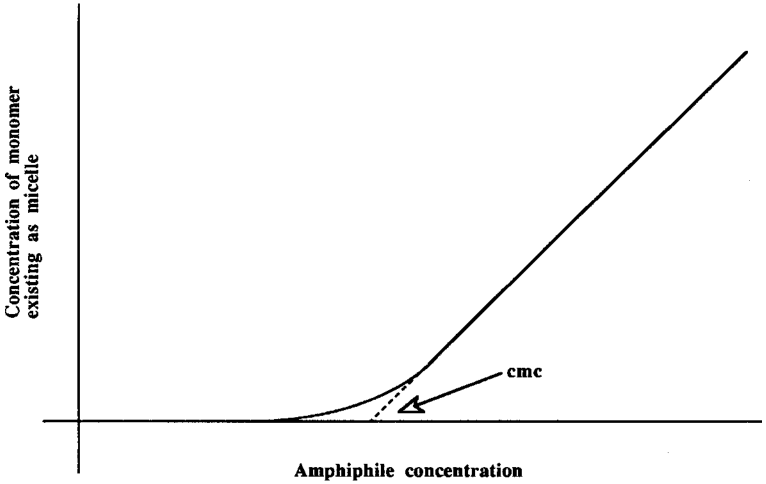


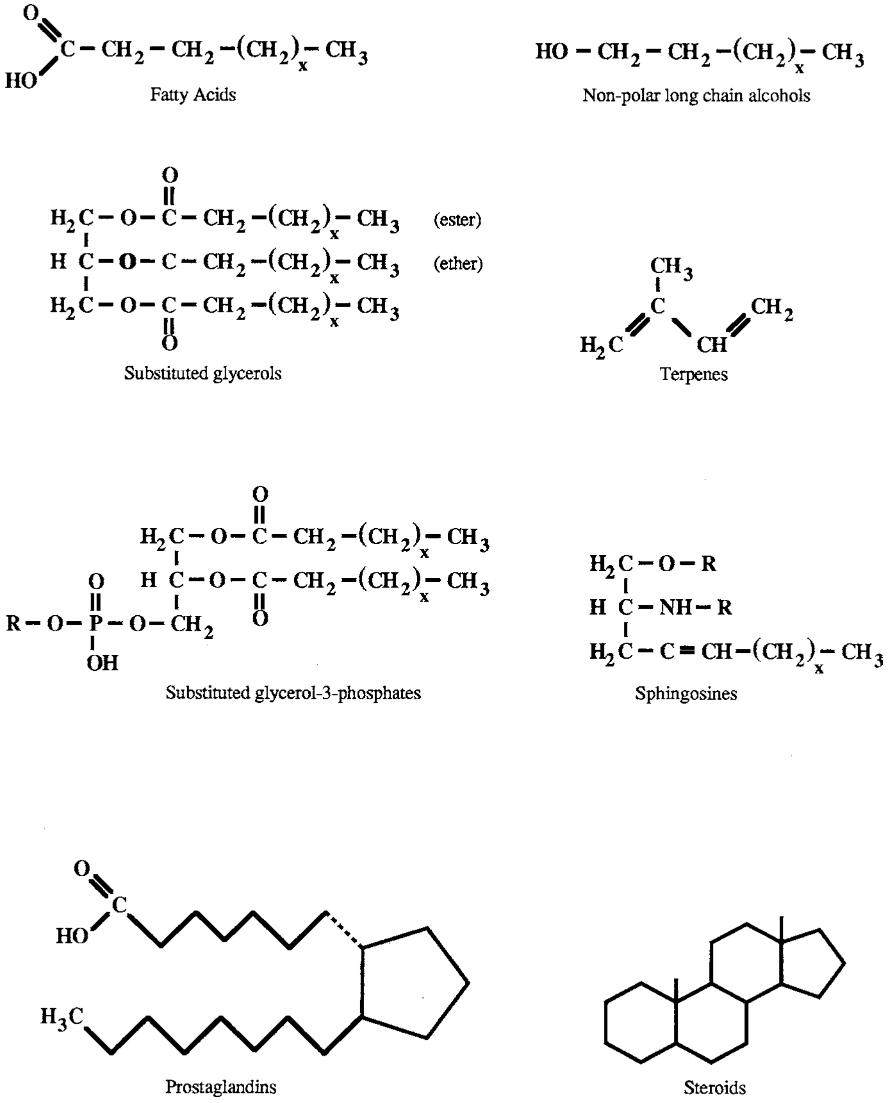
Fig. 17.5 Plot of the incorporation of a solution of amphiphiles into micelles as the concentration of the monomer increases

amphiphile under fixed conditions by varying the concentration alone (Fig. 17.5). The ionic strength and the temperature also are important, since these affect the electrostatic interactions between the polar and "head" groups. The importance of the combined ionic and hydrophobic behavior of large molecules in biological systems cannot be overstated. The search for a systematic understanding of these interactions underlies much of the modern research in biophysical chemistry.

17.4 Lipids Can Be Placed into Several Major Classes

Lipids form a very diverse collection of molecules that can be obtained from cells and tissues by treatment with nonpolar solvents such as chloroform and benzene. They all share the characteristic of being water insoluble, a property derived from the extensive hydrocarbon portion of their structure. Lipids can be classified in a variety of ways. The most generally useful scheme is a classification based on their backbone structure. These structures are listed and illustrated in Fig. 17.6. On the basis of their structure, the families of lipids are considered as fatty acids, nonpolar long-chain alcohols, substituted glycerols, substituted glycerol-3-phosphates, sphingosines, terpenes, prostaglandins, or steroids.

The simplest of the lipids are the single-chain fatty acids. These are large organic ions. Fatty acids are extremely common as components of the more complex lipids found in most cell membranes, although it is uncommon to find them in the free or unesterified form in biological tissues. All fatty acids are composed of a long hydrocarbon chain that terminates in a carboxyl group. At the appropriate pH, the carboxyl group will be ionized, but more commonly the carboxyl group will esterify


O
HO
C – CH2 – CH2 –(CH2)x– CH3
HO – CH2 – CH2 –(CH2)x– CH3
H2C – O – C – CH2 –(CH2)x– CH3
H C – O – C – CH2 –(CH2)x– CH3
H2C – O – C – CH2 –(CH2)x– CH3
(ester)
(ether)
CH3
C
H2C
CH
CH2
R – O – P – O – CH2
OH
H2C – O – R
H C – NH – R
H2C – C = CH –(CH2)x– CH3
H3C

Fatty Acids

Non-polar long chain alcohols

Substituted glycerols

Terpenes

Substituted glycerol-3-phosphates

Sphingosines

Prostaglandins

Steroids

Fig. 17.6 Lipids may be classified on the basis of their backbone structure

with an alcoholic moiety from a suitable backbone such as glycerol to form complex lipids such as the glycerol esters. It is to be noted that a fatty acid is a potential electrolyte, and in the pure state it would be expected to be uncharged. Most fatty acids of biological importance have alkyl chains between 12 and 22 carbons in length. The alkyl chains are either saturated or partially unsaturated. If the hydrocarbon chains contain double bonds, the bonds are almost always found in the *cis*-configuration, although certain unusual fatty acids do contain *trans*-oriented double bonds. The length of the hydrocarbon chain and its degree of unsaturation are important determinants of the physical properties of fatty acids. As Fig. 17.7 indicates, increasing the length of the hydrocarbon chain leads to a rise in melting temperature. The presence of double bonds leads to a significant drop in the melting point when compared to other chains of equivalent length.

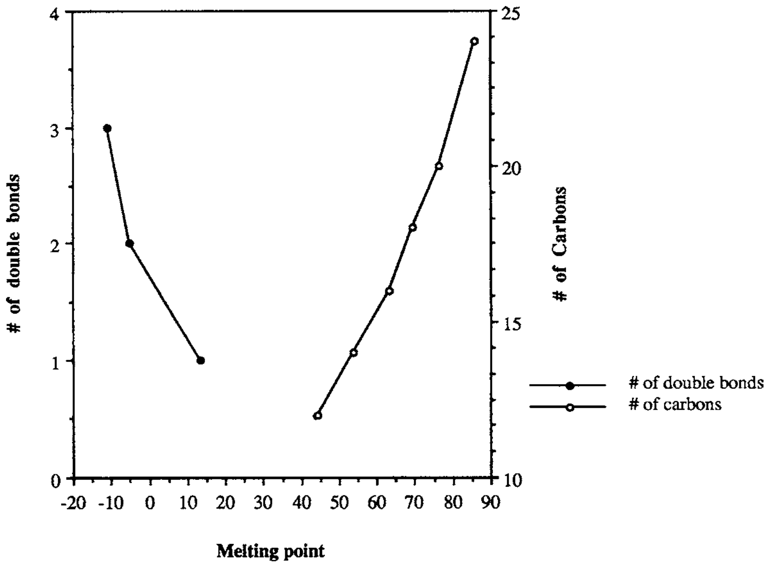


Fig. 17.7 The melting point for fatty acids increases as the chain length increases, as shown by the curve on the *right* representing the melting points for saturated fatty acids as a function of chain length. Conversely, the melting point falls as the number of double bonds increases, as shown by the curve on the *left*, representing the melting points for 18-carbon fatty acids as a function of the number of double bonds

The physical reasons for these variations in thermodynamic properties are easily described. As Fig. 17.8 shows, the saturated hydrocarbon chain has free rotation around all of its carbon–carbon bonds, and hence its minimum energy configuration is a fully extended arrangement of the carbon atoms. Double bonds restrict rotation around the carbons, leading to the formation of rigid kinks in the chains. The maximal hydrophobic interaction between the hydrocarbon chains of the fatty acid is achieved through the lateral interactions. The aliphatic chains form layers interacting side by side with the resultant structures being stabilized by multiple van der Waals interactions. If the aliphatic chains are saturated, the packing can be tighter with an increased number of van der Waals interactions. Conversely, when the aliphatic chains are unsaturated the kinked chains prevent efficient close packing. Unsaturated fatty acids therefore cannot interact as strongly with each other as can the saturated fatty acids. As a result, it takes less heat energy to disrupt the interactions in a solid unsaturated fatty acid compared to a solid saturated fatty acid.

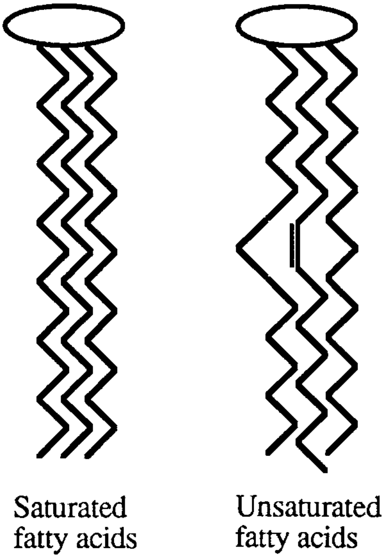


Fig. 17.8 The freedom of rotation around the carbon–carbon bonds in a saturated fatty acid leads to an extended configuration that allows maximal intermolecular interactions to stabilize the structure. When double bonds are present, the chain becomes fixed in *cis* configurations, which prevents efficient packing of the molecules, leading to a less favorable interactional energy

In contrast to these simple single-chain lipids, many of the lipids found in cell membranes are *glycerol esters*. If all three carbons of the glycerol backbone are linked to simple fatty acids via an ester linkage, the compounds are called triacylglycerols, neutral fats, or sometimes triglycerides. These lipids are neutral and somewhat polar due to the ester linkages. As a group, they are extractable in nonpolar solvents. The properties of the glycerol esters depend on the length and degree of unsaturation of the fatty acids attached to each of the three carbons of the glycerol backbone.

If the third carbon of the glycerol backbone is phosphorylated, while the first two each carry a fatty acid ester, the resultant lipid is called *phosphatidic acid*. Because the C-3 position carries a charge, the properties of phosphatidic acid are quite different from those of a neutral fat carrying the same fatty acids on C-1 and C-2. The *phospholipids* are derivatives of phosphatidic acid in which the phosphate group is esterified with an amino acid, an aliphatic amine, a choline residue, a carbohydrate residue, or some other residue (Fig. 17.9).

In mammalian systems, the glycerol ester-based lipids frequently have a saturated fatty acid on C-1 and an unsaturated one on C-2. Like their parent fatty acids, the nonpolar aliphatic substituents of the glycerol esters prefer not to interact with H_2O but will tend to form hydrophobic interactions with each other. If the third carbon has a polar or charged substituent, these moieties will seek stabilization by hydrogen bond and electrostatic interactions. At physiologic pH,

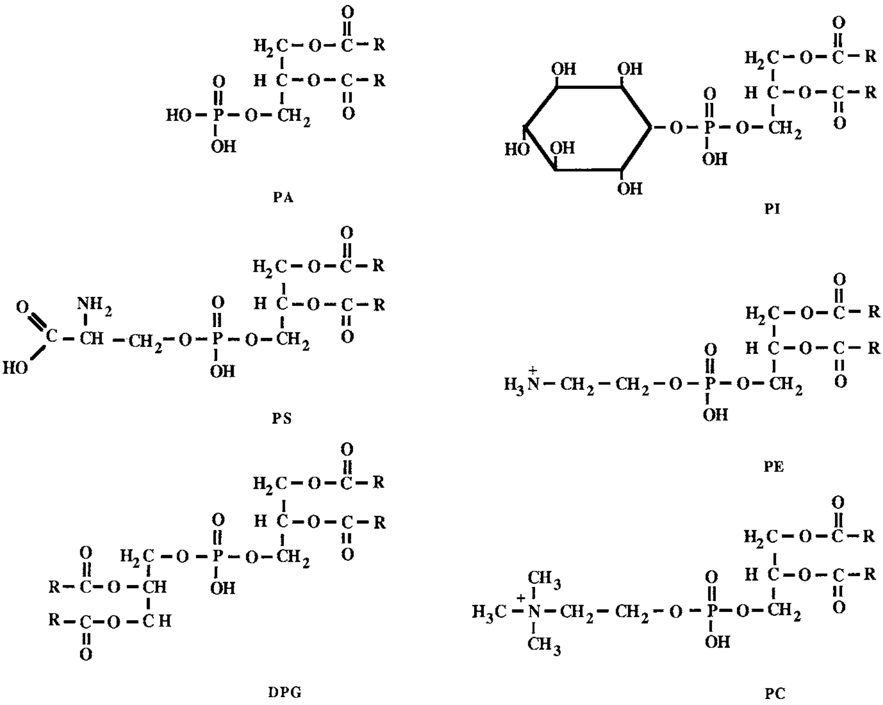


Fig. 17.9 Structures of the major phospholipids found in cell membranes: PA, phosphatidic acid; PS, phosphatidylserine; PC, phosphatidylcholine; PE, phosphatidylethanolamine; PI, phosphatidylinositol; DPG, diphosphatidylglycerol

the phospholipids phosphatidylcholine and phosphatidylethanolamine have no net charge; phosphatidylinositol and phosphatidylserine have a net charge of –1; and phosphatidic acid and diphosphatidylglycerol have a net charge of –2.

Other lipid families are present in membranes, notably the *sphingolipids* and the *sterols*. Sphingolipids are complex lipids that are found in many membranes, but especially in membranes of nerve and brain tissues. All of the sphingolipids contain three components: a long-chain amino alcohol called *sphingosine* or one of its derivatives, a fatty acid, and a polar head group. It is perhaps easier to visualize the structure of the family of sphingolipids by considering the core unit of the molecule to be 1-amino-2-hydroxyethane, which is then substituted as shown in Fig. 17.10. The sphingolipids characteristically have two nonpolar tails and a polar head group. The most abundant sphingolipids are the sphingomyelins, which contain either phosphorylethanolamine or phosphorylcholine. Like the phospholipids phosphatidylserine and phosphatidylethanolamine, these molecules are zwitterions at neutral pH. If the polar head group of a sphingolipid is comprised of a neutral saccharide or polysaccharide, a neutral sphingolipid results respectively called a *cerebroside*or *ganglioside*. It is interesting to note that three gangliosides differing

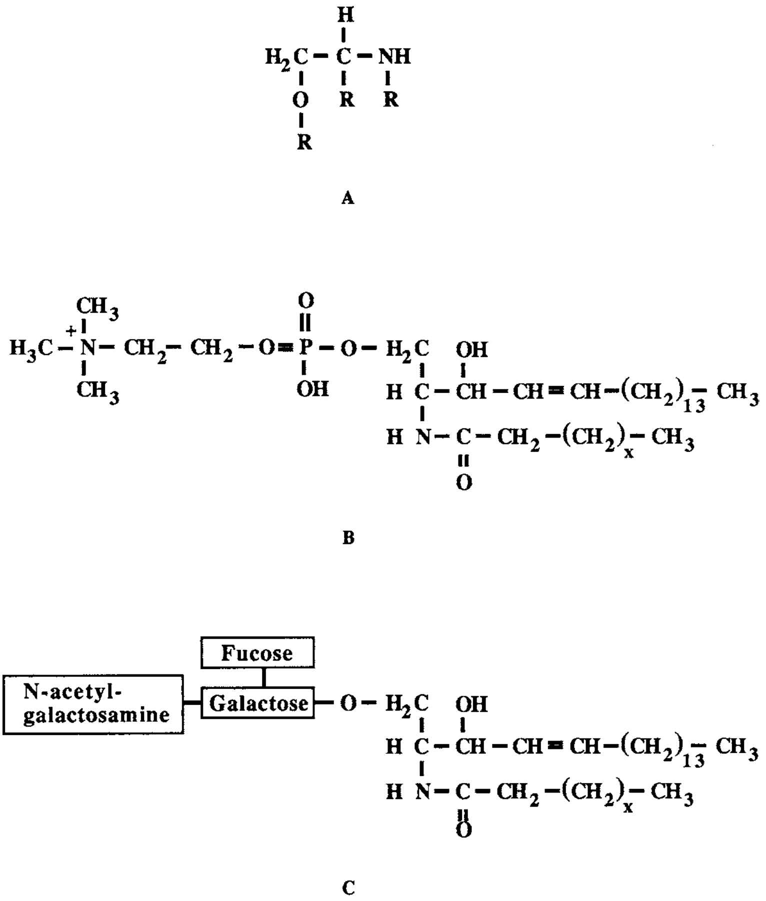
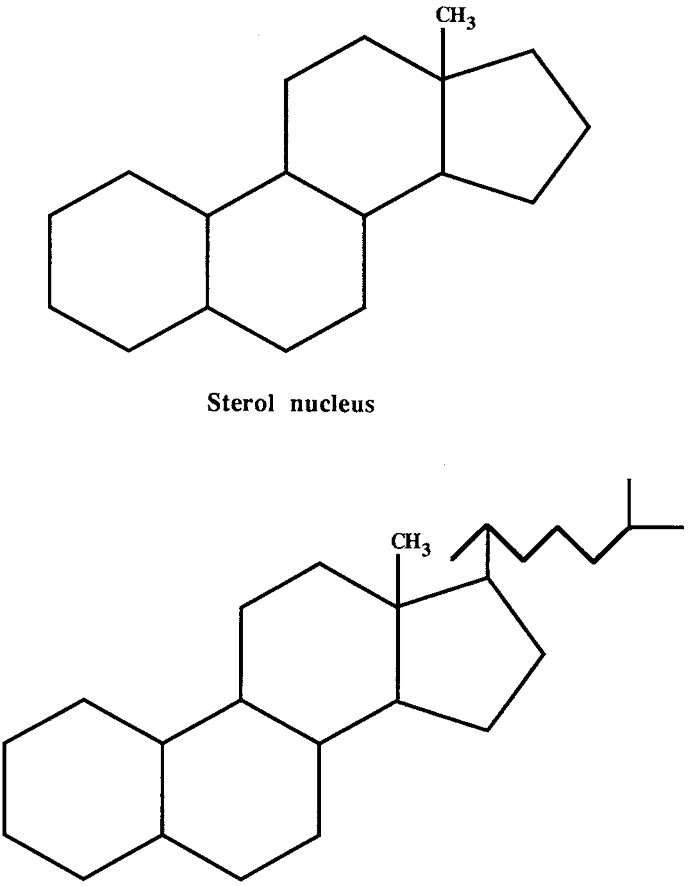


Fig. 17.10 Basic structure of the sphingosine lipids: (**a**) 1-amino-2-hydroxyethane; (**b**) sphingomyelin; (**c**) the ganglioside responsible for the group O antigen of the ABO blood group

only in the polysaccharide structure are responsible for the antigenic determinants of the ABO blood groupings.

The *sterols* are quite distinct as membrane lipids, especially in terms of their molecular structure and shape. Cholesterol is the best known of the animal steroids and is found extensively in cell membranes. Since it is the precursor of many other steroid derivatives, such as bile acids, and hormones, such as androgens, estrogens, and adrenocorticosteroids, it is found widely dispersed in tissues. While the nonpolar chains of the previous classes of lipids were comprised of predominantly single carbon – carbon-bonded aliphatic chains, the nonpolar structure of the steroids is a planar ring structure of significant rigidity (Fig. 17.11). When cholesterol inserts into the lipid bilayer of the membrane, the rigid cholesterol nucleus affects the fluidity of the membrane.



Sterol nucleus

Cholesterol

Fig. 17.11 Structure of the sterol nucleus of the lipid cholesterol. The nonpolar structure of the steroids is a planar ring structure of significant rigidity

17.5 The Organization of Lipids into Membranes Occurs When Aqueous and Lipid Phases Come in Contact

When simple amphiphilic molecules are found in an aqueous environment above the critical micelle concentration, these lipids associate hydrophobically side by side so that the aliphatic chains maximize van der Waals interactions and exclude water. The charged ends interact electrostatically, thereby either attracting or repelling one another. Molecules containing aliphatic chains of 10–15 carbon atoms will form micelles if the lipid solution is strongly agitated (usually sonicated). The aliphatic chains in a micelle are internal, while the charged ends reside on the surface of the sphere, where maximum hydration can be attained. These same lipids, if gently layered onto an aqueous surface, will form a monolayer in which the polar ends stick into the aqueous medium, and the nonpolar aliphatic chains protrude above the solvent (Fig. 17.12). The surface tension of the water will be markedly reduced.

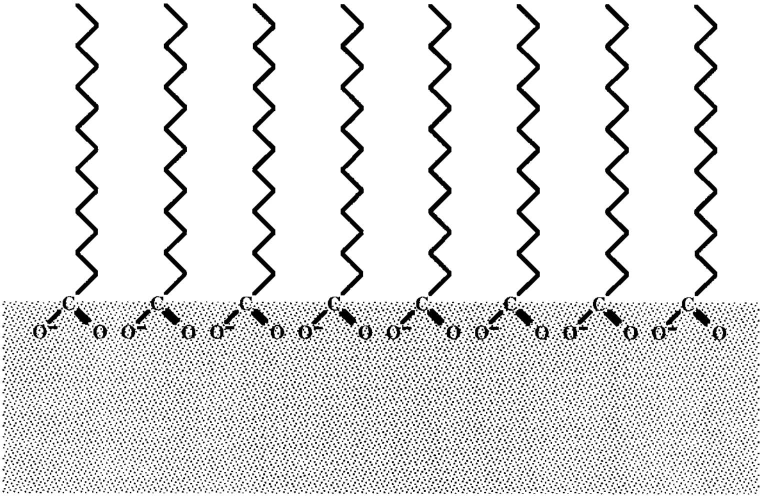


Fig. 17.12 Formation of a monolayer of lipid molecules occurs when a fatty acid is layered onto an aqueous solution. The polar head group associates with the water molecules, and the hydrocarbon tail orients so that it avoids contact with the aqueous solvent

In similar fashion, if a polar complex lipid such as a phospholipid is added to water in a very low concentration, the lipid molecules may be dissolved by inclusion into clathrates. Above its critical micelle concentration, the lipid will associate into a thermodynamically favored aggregation such as a micelle or a lamellar *bilayer*. The formation of these structures is favored by a ΔG of approximately –60 kJ per mole of lipid transferred from water to a micelle. The types of lipid structures resulting from these interactions are shown in Fig. 17.13. The interaction of these polar lipid assemblies with water and the structures that result depend strongly upon the nature of the nonpolar and polar substituents. In every case, these structures will orient so as to expose the polar residues to the aqueous milieu and the nonpolar aliphatic regions to each other. If the lipids are deposited gently at the interface between air and water, a monolayer will be produced with the aliphatic chains protruding from the water and the polar ends interacting with the aqueous solvent at the interface. If the lipid–water system is strongly agitated, the lipids will be dispersed and will form micelles in which a monolayer of lipid forms an array with the aliphatic regions interacting only with each other, while the amount of surface exposure of polar residues to the aqueous solvent is maximized. These structures can be spherical or concave in shape.

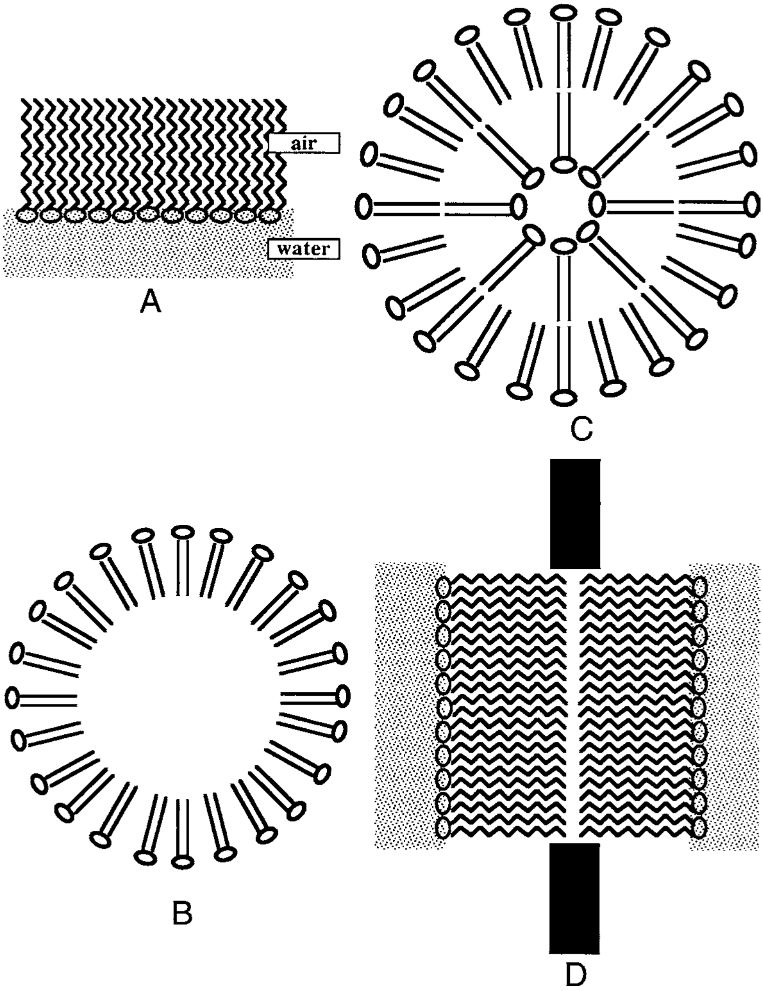


Fig. 17.13 A variety of lipid structures are formed when aggregation of phospholipids in contact with aqueous phases occurs: (**a**) a monolayer formed at an air–liquid interface; (**b**) a micelle; (**c**) the liposome, a bilayer structure, that is self-enclosing and separates two aqueous phases; (**d**) an artificial bilipid membrane, a planar bilayer, that separates two aqueous phases

If a mixture of phospholipids in water aggregates so that water is found both outside and trapped inside the lipid layer, a bilayer structure called a *liposome* will form. In a bilayer structure such as the liposome, a pair of aqueous phases is separated by a lipid phase. The bilayer forms so that the hydrophilic head groups face outward to form an interface with the aqueous phase, while the hydrophobic hydrocarbon chains are protected from contact with water by remaining on the inside of the two leaflets. The nonpolar aliphatic chains on carbons 1 and 2 of the glycerol backbone interact as closely with each other as their degree of unsaturation permits.

The lipid molecules tend to associate side by side so as to maximize both the possible van der Waals interactions between the aliphatic chains and the hydrogen bonds, and electrostatic interactions between the polar ends. The van der Waals energy accounts for a free energy change of approximately –4 to –6 kJ/mol, and the electrostatic and hydrogen bonding interactions account for a free energy change of –4 to –400 kJ/mol, depending on the environment. Furthermore, there is an entropic component favoring the self-association of lipids, since H_2O is excluded from the contact region between the aliphatic nonpolar ends.

Bilayers of this type are considered to be elementary models for biological membranes and have been used extensively in research on membranes. Multiple physical techniques such as nuclear magnetic resonance, electron-spin resonance, electron microscopy, and low-angle x-ray diffraction studies have all confirmed that lipid bilayers of the type described are a fundamental component of biological membranes. Liposomes have been extensively studied not only because they are a reasonable experimental model of cell membranes but also because of their possible use as drug delivery systems for the treatment of disease. These therapeutic studies are directed at utilizing the tendency of lipids to interact with each other. The principle is based on the idea that the bilayer membrane of a liposome will fuse with the cellular plasma membrane, thus mingling its contents with those of the cell and delivering the drugs, probes, or other entities enclosed in the liposome into the cellular cytoplasm. Liposomes, depending on the phospholipid involved, can be internalized inside other larger ones, forming polylamellar or “onionskin” liposomes. If these structures are sonicated, that is, given enough energy to rupture and re-form, the liposomes will be unilamellar, containing a single bilayer of lipids, and will be of uniform size. This size will be dictated by the surface-to-volume ratio, which is dependent on the nature of the lipid.

An important result from the studies of liposomes is that when artificially mixed phospholipid liposomes are formed, there is an asymmetric distribution of the phospholipid types between the inside and outside leaflets. This observation coincides with a variable but demonstrated asymmetry of phospholipid components in the membranes of cells. The asymmetry of the phospholipid distribution is considered to result from a thermodynamically favored configuration in which strain resulting from the curvature of the surface and the electrostatic interactions of the charged head groups is minimized. It is not yet clear if the asymmetry of the membrane components has a physiologic purpose. Furthermore, while the cytoskeleton and the consumption of ATP seem to be necessary for maintenance of the asymmetry, the rationale and details of these observations are not yet known.

Another type of artificial bilayer can be formed, called a *black membrane*. This is a planar synthetic bilayer that is made by introducing the lipids under water at a small hole in a barrier erected between two aqueous compartments. Black membranes have proven to be very useful in modeling mammalian membrane properties, in evaluating the importance of a specific phospholipid and/or of the composition of the C-1 and C-2 fatty acids, and in describing the properties of a membrane. It is clear that the substituents on all three glycerol carbons play a vital role, as do the other membrane-associated molecules such as cholesterol, proteins (including glycoproteins and lipoproteins), and carbohydrates. Among the characteristics of lipids which are important in this role is their structure. In particular, the lipids in mammalian membranes at body temperature are in the *liquid crystalline* form, a partially ordered but nevertheless flexible array which permits the lateral motions demanded by the fluid mosaic model of a mammalian membrane. The membrane lipid structure also affects the conformation of the transmembrane and intrinsic membrane proteins, which in turn affects their function, as receptors, channels, carriers, exchangers, recognition sites, etc. Thus, any modulation of membrane lipid structure or any change of state from one state to another (e.g., crystalline to liquid crystalline) will alter the properties of the membrane and may affect the function of the cell.

17.6 The Physical Properties of Lipid Membranes

17.6.1 Phase Transitions in Lipid Membranes

Lipids undergo changes in state just as other compounds do. A distinguishing characteristic of lipids in membranes is the ability to exhibit an intermediate or mesomorphic state called the *liquid crystalline* or the gel state. A liquid crystalline state is intermediate in level of organization between the rigid crystal and the fluid liquid state (Fig. 17.14). The transition of state between the solid crystalline and the liquid crystalline form in lipid bilayers is sensitive to temperature. In pure lipid bilayers, there is a well-defined melting or *transition temperature*, $\boldsymbol{T}_{\mathrm{m}}$, below which the lipid will be in a solid crystalline phase, and just above which the lipid will usually assume a less ordered liquid crystalline arrangement. In the pure lipid bilayer, there is significant cooperativity in the melting process, which leads to the sharply defined transition temperature. While the discussion in this section will be confined to the simpler cases of single-component membranes, the heterogeneous composition of real membranes does not give rise to a sharp transition point, and melting will occur instead over a several-degree range of temperature.

At a temperature below the melting point, the solid crystalline phase exists in which the nonpolar hydrocarbon tails are rigidly held in an all-*trans* configuration; there is little lateral mobility of each molecule. The all-*trans* arrangement of the nonpolar tails leads to occupation of a minimum volume by the hydrocarbon groups in the membrane. The *cis* configuration that results from the presence of unsaturated bonds in the hydrocarbon tails leads to a greater occupied volume. Because the interactions between the hydrocarbon chains in the *cis* configuration are more limited, the transition temperature is lower than that of chains with an all-*trans* configuration. Below the transition temperature, a saturated hydrocarbon chain (normally in the all-*trans* configuration) will have a very low frequency of *kink* or *cis*-conformation formation, about one kink per 10 acyl chains. As the temperature of the system is increased, there is increasing disturbance of the tightly packed crystalline structure until the transition temperature is reached, where there will be a kink frequency of

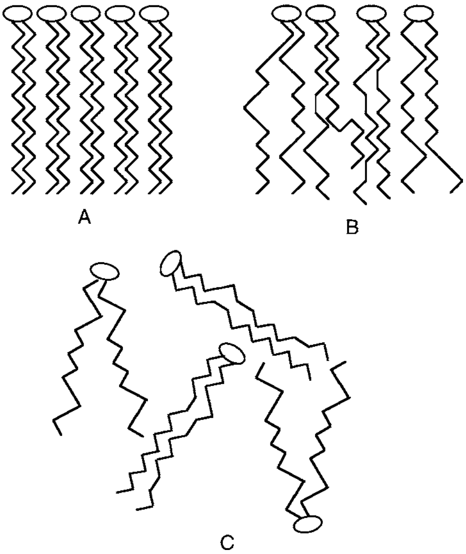


Fig. 17.14 Comparison of the lipid configurations in the (**a**) solid crystalline; (**b**) liquid crystalline; (**c**) liquid states

about one per acyl chain. Above the melting point, the acyl chains will have two or more kinks per chain. The liquid crystalline phase is characterized by a fairly rigid structure near the polar (charged) head groups and a considerably more disorganized and flexible region near the central hydrophobic portion of the bilayer. The loosened organization of the liquid crystal allows for lateral diffusion and a freedom of rotation along the long axis of the hydrocarbon chain. The transition from a stretched all-*trans* conformation to a kinked conformation is accompanied by an increase in membrane volume and a decrease in membrane thickness.

17.6.2 There Are Specific and Limited Motions and Mobilities Found in Membranes

The existence of the liquid crystalline state allows a significant increase in the intra- and inter-molecular movement of components in the membranes. The phase transition to the liquid crystalline state in membranes affects the ordering and fluidity of the membrane. Measures of these properties are of value when describing membrane behavior. The fluidity is described by a term that refers to the viscosity of the membrane. Viscosity is actually a term appropriately used in macroscopic homogeneous phases without anisotropy. Membranes even of pure lipids are by definition anisotropic and non-homogeneous, so the concept of local viscosity or *microviscosity* is used. The microviscosity of a membrane or of a lipid bilayer is a measure of the ease with which lateral motion of lipid molecules with respect to one another can occur in that membrane or bilayer. The microviscosity of a membrane in the liquid crystalline phase is highest near its surface and lowest in its center.

This is true whether the bilayer contains pure lipids, as is true in liposomes, or whether proteins, glycoproteins, and lipoproteins have been inserted, as in a typical mammalian membrane. This property is important in the fluid mosaic model of a mammalian membrane because it contributes to the velocity with which transmembrane proteins can travel toward each other in order to form a membrane protein assembly (a “cap”) and, thus, may control the energy of activation required for the conformational changes occurring in the membrane, such as those associated with receptor-mediated functions. The microviscosity is a function of the liquid crystalline structure of the membrane, and the energy required to achieve a transition from one liquid crystalline state to another as well as the temperature at which such a transition occurs can be calculated from the thermal denaturation curves of a membrane in a fashion similar to that to be described for macromolecules in Chapter 18. The microviscosity is sensitive to temperature, to pressure applied to the membrane, and to the chemical composition of the membrane.

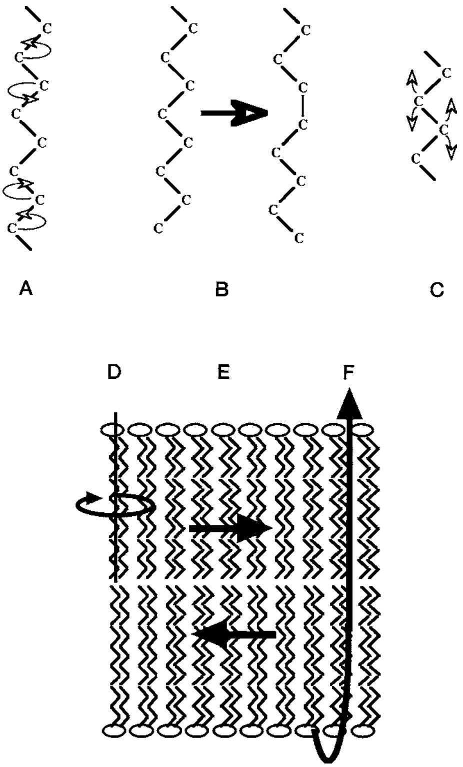
The microviscosity can be evaluated by various techniques such as

(1) ^{13}C nuclear magnetic resonance (NMR) to measure the freedom of motion around C–C bonds.
(2) Electron paramagnetic resonance (EPR) to measure the freedom of rotation of a spin probe, usually introduced into the membrane as a nitroxide attached at the end of a long aliphatic chain.
(3) Fluorescence polarization techniques to evaluate, via depolarization of the incident polarized light beam, the extent to which a lipophilic probe’s fluorescence emission is altered.

Each of these techniques has limitations. NMR requires a high concentration of phospholipid and is consequently difficult to apply to cellular systems unless the cells tolerate relatively tight packing. Furthermore, NMR cannot detect carbon–carbon movement whose frequency falls below 10^5 s^{-1}. Since the paramagnetic tag covalently linked to an EPR probe cannot penetrate the membrane, there is a risk that EPR will measure the fluidity of, at best, the outer surface of the cell membrane and, at worst, the spinning of the probe itself outside the membrane. If fluorescence depolarization is used, the choice of probe dictates the region that is being examined, since some probes are completely embedded in the membrane while others protrude into the extracellular milieu because they are charged. Further details of the actual measurement of microviscosity are beyond the scope of this text and will not be discussed. It suffices for the reader to know that the fluidity of a membrane can be correlated with its phospholipid composition, with the degree to which the fatty acid components of the phospholipids are unsaturated, and with the level of incorporation of intercalating entities, such as cholesterol, in the membrane.

A variety of modes of motion can occur in lipid membranes (Fig. 17.15). The most common of these is rotation around the carbon–carbon bonds in the long axis of the acyl chain. The speed of these rotational events is on the order of 100 ps. The lifetime of the *cis* kinks in the hydrocarbon chains is about 10 times longer, about 1 ns. The lipid molecules also undergo bending motions. With the exception of the

Fig. 17.15 Various modes of motion found for lipids in membranes: (**a**) intra-chain rotation around carbon–carbon bonds; (**b**) kink formation; (**c**) bending of the hydrocarbon chain; (**d**) rotational diffusion; (**e**) lateral diffusion; (**f**) transmembrane flip-flop movement



formation of kinks, which is tied to the fluidity of the membrane, these carbon–carbon rotations and bends seem to have no clear functional relevance. Rotation of the entire lipid molecule including the head group also occurs in the membrane with a rotational speed on the order of 2π radians/10 ns. The parameter describing this rotational motion is called *rotational diffusion*.

Perhaps one of the most important motions that will occur in a membrane is *lateral diffusion*. Lateral diffusion can be thought of in terms similar to those that are the basis of the treatment of Brownian motion by rate theory, which will be covered in some mathematical detail in Chapter 22. Lateral diffusion can be thought of as a process where a lipid molecule will experience net translational motion in the plane of the membrane by executing a series of jumps from one vacancy in the liquid crystalline array to another. The rate of lateral jumping in membranes is on the order of 10^7–10^8 s^{-1}, while the net movement is much less, on the order of micrometers per second. The rates of lateral diffusion are strongly affected by the presence of membrane-associated proteins and by immobilized lipid molecules closely associated with these membrane proteins.

The form of motion with the longest time frame is the *transmembrane flip-flop* of a lipid molecule. This motion requires the complete translocation of a lipid molecule from one half of the bilayer to the other leaflet. The fastest rate of exchange is reported to be in minutes, but frequently the equilibration times for flip-flopping are in the range of hours to weeks. The slowness of the process can be explained by the enormous energy barrier posed by the required movement of the polar head group through the hydrophobic interior of the bilayer in the course of the translocation. The kinetics of the flip-flopping can be shortened by the presence of certain membrane proteins, just as other parameters of membrane motion were affected by the nonhomogeneous state of the biological membrane.

17.7 Biological Membranes: A More Complete Picture

A biological membrane acts as a separator between two solutions, permeable to some but not to all components of these solutions. In mammalian cells, several types of membranes exist, including the plasma membrane separating the cytoplasm from the extracellular milieu and different organellar membranes separating the cytoplasm from the inner contents of organelles such as mitochondria or secretory granules. The modern model of the biological membrane is called the *fluid mosaic model* and is described as a lipid bilayer with an array of various proteins “floating” in the bilayer structure (Fig. 17.16).

These membranes are phospholipid bilayers which contain membrane-spanning proteins and also, in many cases, membrane-bound proteins which do not traverse the membrane. In the case of the plasma membrane separating the cytoplasm from the exterior of the cell, the external portion of the proteins may be glycosylated. The membrane proteins act as carriers, pumps, channels, and, in general, mediators of

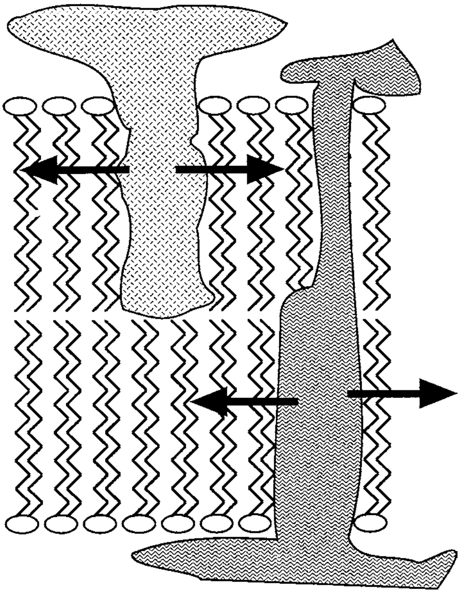


Fig. 17.16 Fluid mosaic model of the cell membrane

cell function and/or sites of cell–cell interactions or cell recognition. In all of these capacities, the conformation of the protein plays a critical role. That conformation is dependent on the interactions between the protein and the membrane with which it comes in contact and, therefore, on the structure of the membrane. That structure, as discussed in this chapter, depends on the nature of the lipids, which also controls the microviscosity, the transition temperatures of the membrane lipids, and also the ease of passage of water, ions, and uncharged entities from one side of the membrane to the other. Some proteins project into the cytoplasmic space, while other proteins are only inserted into one side or bound to one side of the membrane. Many intrinsic membrane proteins traverse the membrane several times (e.g., seven times for the β-adrenergic receptors and rhodopsin, the main visual protein). Each such traversing region contains a sequence of hydrophobic residues capable of forming a helix. The role of each exposure on the exterior or cytoplasmic sides of the membrane is not yet clear.

Further Reading

General

Israelachvili J. (1992) *Intermolecular and Surface Forces*, 2nd edition. Academic Press, London.

Kotyk A., Janacek K., and Koryta J. (1988) *Biophysical Chemistry of Membrane Functions*. Wiley, New York.

Nagle J.F. (1980) Theory of the main lipid bilayer phase transition, *Annu. Rev. Phys. Chem.*, **31**:157–195.

Small D.M. (ed.) (1986) *The Physical Chemistry of Lipids*. Handbook of Lipid Research, Volume 4. Plenum Press, New York.

Solute–Solvent Interaction

More details on aqueous clathrate structure, both the theories and the controversies, can be found in the series edited by Franks and other listings in Chapter 14.

Articles on Clathrate Organization Around Biomolecules

Byrn M.P., Curtis C.J., Hsiou Y., Khan S.I., Sawin P.A., Tendick S.K., Terzis A., and Strouse C.E. (1993) Porphyrin sponges: conservation of host structure in over 200 porphyrin-based lattice clathrates, *J. Am. Chem. Soc.*, **115**:9480–9497. (This article discusses the application of the clathrate ideas to a non-aqueous highly structured “solvent,” tetraarylporphyrin molecules!)

Hvidt A. (1983) Interactions of water with nonpolar solutes, *Ann Rev. Biophys. Bioeng*., **12**:1–20.

Lipscomb L.A., Peek M.E., Zhou F.X., Bertrand J.A., VanDerveer D., and Williams L.D. (1994) Water ring structure at DNA interfaces: hydration and dynamics of DNA-anthracycline complexes. *Biochemistry*, **33**:3649–3659.

Lipscomb L.A., Zhou F.X., and Williams L.D. (1996) Clathrate hydrates are poor models of biomolecule hydration. *Biopolymers*, **38**:177–181.

Teeter M.M. (1992) Order and disorder in water structure of crystalline proteins. *Proc. Natl. Acad. Sci. USA.*, **74**:63–72.

Wolfenden R. and Radzicka A (1994) On the probability of finding a water molecule in a nonpolar cavity, Science, **265**:936–937.

Artificial Membranes

Ostro M.J. (1987) Liposomes, *Sci. Am.*, **256, 1**:102–111.

Schindler H. (1989) Planar lipid-protein membranes: strategies of formation and of detecting dependencies of ion transport functions on membrane conditions, *Methods Enzymol.*, **171**:225–253. (Oriented to laboratory work. A relatively recent review of the field of artificial membranes.)

Biological Membranes

Gennix R.B. (1989) *Biomembranes: Molecular Structure and Function.* Springer-Verlag, New York.

Kohlwein S.D (1992) Biological membranes, function and assembly, *J. Chem. Educ.*, **69**:3–9.

Sharon N. and Lis H. (1993) Carbohydrates in cell recognition, *Sci. Am.*, **268, 1**:82–89.

Singer S.J. and Nicholson G.L. (1972) The fluid mosaic model of the structure of cell membranes, *Science*, **175**:720–731.

Chapter 18
Macromolecules in Solution

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_18, © Springer Science+Business Media, LLC 2010

18.1 The Physical Interactions of Polymers in Solution Are Not Unique but Modeling the Interactions Will Require Different Considerations Than Those of Smaller Molecules

The solvent–solute structure in aqueous solutions of small polar and nonpolar molecules has been described. The view presented of the disturbance of the normal hexagonal structure of liquid water by these small solutes can serve as a model for solutions of macromolecules, which, to a reasonable approximation, may be treated as a chain of small solute molecules attached to each other. Such molecules are *polymers.* Polymers of a single component or monomer are called *homopolymers* (A_n). *Copolymers* are comprised of two or more monomers. Copolymers can be random (AAABBABBBA...) or ordered (ABABABAB or ABCABCABC). Polymers of these types are mainstays of the chemical industry, and their solubility or insolubility in aqueous environments can now be predicted with a high degree of accuracy, allowing proper formulation of the polymerizing mixture.

Ultimately, we are interested in biological systems where the polymers of interest are not the synthetic plastics and fabrics, but biological polymers such as proteins (polyamino acids), polynucleotides, glycopolymers (polysugars), or mixed biological polymers such as proteoglycans. When a portion of such a polymer is comprised of some charged residues, a *polyelectrolyte* results. Few naturally occurring proteinaceous biological polymers consist of a single component (e.g., AAAAAAA), although cellulose and glycogen are polymers of the single hexose glucose. Thus, most polymers of biological origin are heteropolymers whose structure and solvation must be approximated from simpler model systems.

Many biological polymers are rigid. Other polymers are elastic like rubber or the protein elastin. Finally polymers may be viscoelastic with the quality of being gooey or slimy like a raw egg. A primary goal of biophysical chemistry is to define the configuration of a macromolecule. A closely related goal is to be able to define the rules that will allow accurate prediction of the structure and properties of those macromolecules such as its rigidity, elasticity, and viscosity. As can be easily imagined, the canon of chain configurations of even the moderately sized polymer can be substantial. If a heteropolymer is composed of 100 residues each of which may take just one of three configurations, the number of possible conformations of the polymer is 3^{100} or 5×10^{47}. Fortunately, biopolymers represent a special class of polymers in which there are substantial constraints that serve, in many situations, to limit the size of the canonical distribution. We will explore how a statistical modeling process can provide insights into a polymer's characterization first by examining the thermodynamics and then some aspects of the configuration of polymers. We will be able to apply statistical methods with which we have already had some experience.

18.2 Thermodynamics of Solutions of Polymers

In general polymers are derived from the covalent linking of monomers leading to either a linear chain or a branching chain structure. Polymers have distributions of either varying chain length or varying conformations. There are many conformational degrees of freedom and the ability of the polymer to sample these conformations is reflected in the entropy of a polymer and seen in properties such as elasticity of rubber and the gooey viscoelasticity of latex glue. If the various rotational isomers of a chain have nearly the same energy, multiple conformations of the polymer will occur. The properties of a polymer will be greatly influenced by the number of these conformations that can occur at different temperatures.

The thermodynamic properties of polymer solutions are quite different from the solutions of small molecules that we have examined up to this point. These differences are that

- while small molecule solutions have a strong entropy of mixing, polymeric solutions do not;
- the enthalpy of mixing of a polymer and solvent cannot be zero because there will always be attractive forces between the macromolecule and the solvent and also between the polymer segments themselves;
- the colligative properties such as the vapor pressure of a polymer in a small molecule solvent is much lower than that of the same solvent in a small molecule solvent solution.

These differences occur because the polymer molecule is much larger than the solvent molecule. While a single monomer unit may be close to the size of a solvent molecule, a polymer chain may be made from hundreds to thousands of monomers. Thus the volume occupied by a polymer will be very much larger than that of a single monomer. The result is that the mole fraction of a component in these solutions will be very different from the volume fraction. In our development of a theoretical treatment of the thermodynamics of macromolecular solutions (below) we will see that the *partial specific volume* or *volume fraction* is a better measure of polymer concentration when compared to the mole fraction used for small molecule solutions.

We will now explore aspects of a model of polymer–solvent interaction developed by P.J. Flory and independently by M.L. Huggins called the *Flory–Huggins model*. This model makes the assumption that a *lattice description* (similar to that used to describe the crystal structure of a molecule) can be used to characterize the change in molecular conformation of a polymer in solution with solvent molecules. The lattice is considered to be three-dimensional, composed of sites or holes with specific spatial dimension defined as the size of the solvent molecule. The total number of sites is $\boldsymbol{M}$. Each site in the lattice will be filled with either a solvent molecule or a polymer segment. The question to be answered is, In how many different ways can the lattice be filled? This model is a spatial model and each solvent molecule takes up the same space as a polymer segment. In the Flory model the segment is not necessarily the same as the monomer itself but is defined as “the portion of a polymer molecule requiring the same space as a molecule of solvent.” In addition an important assumption is made that the interaction energies that include the monomer–monomer and monomer–solvent interactions are ignored. The Flory model differs from the treatment of a small molecule ideal solution in which the solute and solvent are viewed as the same size and thus differentiable by number, i.e., partial molar concentration. In this model, each polymer molecule is not counted equivalent to each solvent molecule but rather each polymer segment (as defined above) is counted equivalent to a solvent molecule. A polymer may be described by the number of segments in the chain *y*; *y* will be ratio of molar volume of solute molecules to solvent.

Consider the simplest case of a binary solution. A number of solvent molecules, n_S, and a number of polymer molecules, n_P, are mixed to form a solution. Each solvent molecule and polymer segment must be placed in an available site. Solvent molecules can be placed in any available site but the polymer segments have a much more limited choice. Each segment can only be placed into a site in which there is also an adjacent site available for the next polymer segment. A given site will have a certain number of nearest or first-neighbor sites. The number of these first-neighbor coordination sites is given by the variable *z*. Depending on the three-dimensional structure of the lattice, *z* will have a magnitude varying from 6 (for a regular cubic lattice) to 12 (for a hexagonal lattice). Polymer and solvent are added to the lattice and some random arrangement is found (Fig. 18.1). If the lattice is completely filled, then

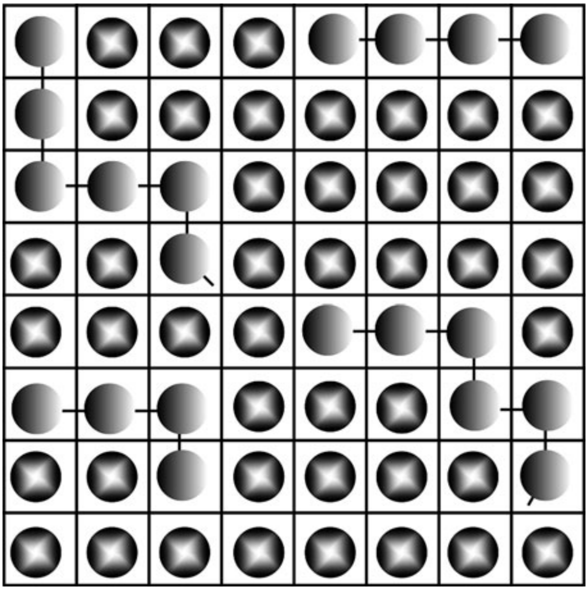


Fig. 18.1 Lattice model for polymer and solvent interaction

$$\boldsymbol{M} = n_{\mathrm{s}} + y\, n_{\mathrm{p}} \tag{18.1}$$

The molar and volume fractions of polymer and solvent are molar fraction volume fraction

$$\begin{array}{lll} \text{Polymer} & \chi_{\mathrm{p}} = \dfrac{n_{\mathrm{p}}}{(n_{\mathrm{p}} + n_{\mathrm{s}})} & \phi_{\mathrm{p}} = \dfrac{y n_{\mathrm{p}}}{M} \\ \text{Solute} & \chi_{\mathrm{s}} = \dfrac{n_{\mathrm{s}}}{(n_{\mathrm{p}} + n_{\mathrm{s}})} & \phi_{\mathrm{s}} = \dfrac{n_{\mathrm{s}}}{M} \end{array} \tag{18.2}$$

18.2.1 The Entropy of Mixing for a Polymer Solution Requires a Statistical Approach

The approach to determining entropy is statistical mechanical in nature and we will seek to express S in terms of the Boltzmann–Planck equation that was derived in Chapter 11:

$$S = k \ \ln \ W \tag{10.67}$$

The entropy of mixing can be found by counting the number of arrangements of n_{p} polymers and n_{s} solvent molecules within the lattice. The starting point is with a lattice that has a certain number of polymer molecules, i, already inserted. The next polymer (i + 1) must now be added. This can best be done by the segment-by-segment addition of the i + 1 polymer. There are initially $M – yi$ vacant cells and once the first polymer segment is added the second segment will face the possibility that a nearest-neighbor site will not be available. There is a probability, f_i, that the site adjacent is occupied. The probability that a site will be available is given by $(1 – f_i)$. What is the number of ways, v_{i+1}, that the i + 1 polymer could fill the cell sites? The answer will depend on the product of the number of ways that the chain of lattice sites can be filled. These can be listed as follows:

$$\begin{array}{rl} \text{for the first site :} & M - yi \\ \text{for the second:} & z(1 - f_i) \\ \text{for the third:} & (z - 1)(1 - f_i) \\ \text{for the fourth:} & (z - 1)(1 - f_i) \\ \text{for the} \quad (y - 2)\text{th :} & (z - 1)(1 - f_i) \end{array} \tag{18.3}$$

Combined, this is

$$v_i + 1 = [M - yi][z(1 - f_i)][(z - 1)(1 - f_i)][(z - 1)(1 - f_i)] \cdots [(z - 1)(1 - f_i)] \tag{18.4}$$

This simplifies to

$$v_{i+1} = (M - yi)(z)(z - 1)^{y-2}(1 - f_i)^{y-1} \tag{18.5}$$

18.2.1.1 Polymer Segments Are Not Distinguishable and Must Be Treated Using the Correct Statistical Description

Equation (18.5) is appropriate if each segment is distinguishable but this is not the case. If polymer segments were exchanged there would be no way to tell them apart; therefore Eq. (18.5) should be corrected to account for the indistinguishable nature of the polymer segments (vide, Chapter 10). An expression must be written that allows for the consideration of all of the configurations for all n_p molecules in the lattice:

$$W_\mathrm{config} = \frac{1}{n_\mathrm{p}!}\prod_{i=1}^{n_\mathrm{p}} v_i = \frac{1}{n_\mathrm{p}!}\prod_{i=0}^{n_\mathrm{p}-1} v_{i+1} \tag{18.6}$$

We must now combine the thinking that led to the term v_{i+1} in Eq. (18.5) with the indistinguishable nature of the polymer segments captured in Eq. (18.6).

The term $(1 - f_i)$ can be determined by asking, What is the average probability, $\overline{f_i}$, that a site is unavailable? The following approximation is made:

$$1 - f_i \approx 1 - \overline{f_i} = \frac{\text{Number of sites available}}{\text{Total number of sites}} = \frac{M - yi}{M} \tag{18.7}$$

We can rewrite Eq. (18.5):

$$v_{i+1} = (M - yi)(z)(z - 1)^{y-2}\left(\frac{M - yi}{M}\right)^{y-1} \tag{18.8}$$

The lone z in Eq. (18.8) can be reasonably replaced by $(z - 1)$ and with a little algebra Eq. (18.8) can be simplified:

$$v_{i+1} = (M - yi)^y\left(\frac{z - 1}{M}\right)^{y-1} \tag{18.9}$$

The first term $(M - yi)^y$ can also be approximated as follows:

$$(M - yi)^y \cong \frac{(M - yi)!}{\left[M - y\,(1 + 1)\right]!} \tag{18.10}$$

Combining Eqs. (18.9) and (18.10) yields

$$v_{i+1} = \frac{(M - yi)!}{\left[M - y\,(1 + 1)\right]!}\left(\frac{z - 1}{M}\right)^{y-1} \tag{18.11}$$

Now W, the number of arrangements for polymers made from indistinguishable polymer segments can be written as

$$W_{\text{config}} = \frac{1}{n_{\text{p}}!} \prod_{i=0}^{n_{\text{p}}-1} v_{i+1} \tag{18.6}$$

$$\begin{aligned} W_{\text{config}} = \frac{1}{n_{\text{p}}!} \Bigg\{ & \underbrace{\left(\frac{M!}{(M-y)!}\right)}_{\text{this is } i=0} \underbrace{\left(\frac{(M-y)!}{(M-2y)!}\right)}_{\text{this is } i=1} \underbrace{\left(\frac{(M-2y)!}{(M-3y)!}\right)}_{\text{this is } i=2} \\ & \ldots \underbrace{\left(\frac{\left(M-\left(n_{\text{p}}-1\right)y\right)!}{\left(M-n_{\text{p}}y\right)!}\right)}_{\text{this is } i=n_p-1} \Bigg\} \left(\frac{z-1}{M}\right)^{n_{\text{p}}(y-1)} \end{aligned} \tag{18.12}$$

$$W_{\text{config}} = \left\{ \frac{1}{n_{\text{p}}!} \frac{M!}{\left(M-n_{\text{p}}y\right)!} \right\} \left(\frac{z-1}{M}\right)^{n_{\text{p}}(y-1)} \tag{18.13}$$

Now we have the expression that we need to evaluate in terms of the Boltzmann–Planck relation:

$$S_{\text{config}} = k \ln W_{\text{config}} \tag{18.14}$$

Combining these now gives

$$S_{\text{config}} = k \ln \left(\frac{1}{n_{\text{p}}!} \frac{M!}{\left(M-n_{\text{p}}y\right)!} \left(\frac{z-1}{M}\right)^{n_{\text{p}}(y-1)} \right) \tag{18.15}$$

We will now mix a lattice of pure solvent (no polymer, n_{s} solvent) and a lattice of pure polymer (no solvent n_{p} polymer) together to get a lattice of the binary mixture (n_{s}, n_{p}). There will be a Boltzmann relationship for the ΔS_{mixing} for this similar to Eq. (10.68):

$$\frac{\Delta S}{k} = \ln \frac{W_2}{W_1} = \ln \frac{W_{\text{mixed}}}{W_{\text{initial}}} = \ln \frac{W(n_{\text{s}}, n_{\text{p}})}{W\left(n_{\text{s}}, 0\right), \left(0, n_{\text{p}}\right)} \tag{18.16}$$

For the lattice of pure solvent

$$W\left(n_{\text{s}}, 0\right) = 1 \tag{18.17}$$

For pure polymer, use Eq. (18.13) and because there are zero solvent molecules we substitute $n_{\text{p}}y$ for M and simplify

$$W\left(0, n_{\text{p}}\right) = \left\{ \frac{1}{n_{\text{p}}!} \frac{n_{\text{p}}y!}{(0)!} \right\} \left(\frac{z-1}{n_{\text{p}}y}\right)^{n_{\text{p}}(y-1)} \tag{18.18}$$

For the mixture, $W(n_s, n_p)$ we add in n_s to Eq. (18.13) by letting $n_s = M - n_p y$:

$$W(n_s, n_p) = \left\{ \frac{1}{n_p!} \frac{M!}{(n_s)!} \right\} \left(\frac{z-1}{M} \right)^{n_p(y-1)} \tag{18.19}$$

Now the entire Boltzmann relationship can be written as

$$\frac{W(n_s, n_p)}{W\,(n_s, 0)\,, \left(0, n_p\right)} = \frac{\left\{ \frac{1}{n_p!} \frac{M!}{(n_s)!} \right\} \left(\frac{z-1}{M} \right)^{n_p(y-1)}}{\left\{ \frac{1}{n_p!} \frac{n_p y!}{(0)!} \right\} \left(\frac{z-1}{n_p y} \right)^{n_p(y-1)}} \tag{18.20}$$

$$= \left(\frac{M!}{\left(y n_p\right)! n_s!} \right) \left(\frac{y n_p}{M} \right)^{n_p(y-1)}$$

Earlier we have discussed the use of the Stirling approximation ($\ln n! = n \ln n - n$) to compute the factorial expressions found in Eq. (18.16). Using this and some algebraic manipulation gives the following result:

$$\frac{\Delta S_{\text{mixing}}}{k} = M \ln M - \left(y n_p\right) \ln \left(y n_p\right) - n_s \ln n_s + n_p\,(y-1) \ln \left(\frac{y n_p}{M} \right) \tag{18.21}$$

Now we manipulate this result using Eq. (18.1) to arrive at an expression for the entropy of mixing in terms of the volume fractions (Eq. (18.2)):

$$\begin{aligned} \frac{\Delta S_{\text{mix}}}{k} =& \left(n_s + y n_p\right) \ln M - n_s \ln n_s - y n_p \ln \left(y n_p\right) + y n_p \\ & + y n_p \ln \left(\frac{y n_p}{M} \right) - n_p \ln \left(\frac{y n_p}{M} \right) \\ =& -n_s \ln \left(\frac{n_s}{M} \right) - y n_p \ln \left(\frac{y n_p}{M} \right) + y n_p \ln \left(\frac{y n_p}{M} \right) - n_p \ln \left(\frac{y n_p}{M} \right) \\ =& -n_s \ln \left(\frac{n_s}{M} \right) - n_p \ln \left(\frac{y n_p}{M} \right) \\ =& -n_s \ln \phi_s - n_p \ln \phi_p \end{aligned} \tag{18.22}$$

This is the expression for the entropy of mixing that we have been seeking. It can be compared to the van't Hoff equation for the ΔS_{mix} for an ideal solution that is written in mole fraction:

$$\Delta S_m = -k \left(n_s \ln \chi_1 + n_p \ln \chi_2\right) \tag{18.23}$$

The volume fraction of Eq. (18.23) accounts for the large molecular interactions that must be included in the description of polymer solutions.

18.2.2 The Enthalpy of Mixing in a Polymer Solution Is Dominated by van der Waals Interactions

For an ideal solution of small molecules in solvent, $\Delta H_{\text{mix}} = 0$. This will not be the case for a polymer solution because there will be interactions between (1) the solvent molecules and the macromolecular chain and (2) between the polymer segments within the chain itself. There are no surprising new physical concepts that underlie these interactions: as discussed in the previous chapters, electrostatic forces form their basis. In the case of polyelectrolytes, explicit ion–ion and ion–dipole-type interactions will be dominant and in the case of polyhydroxylated polymers hydrogen bonding plays a substantial role. However, in the case of many polymers the dominant interactions are of the van der Waals type. The three component factors of these short range forces have been detailed in Chapter 8 and are restated here for review: (1) permanent dipole–permanent dipole, (2) induced dipole–dipole interactions, and (3) dispersion effects. Historically there have been two approaches to the ΔH of mixing for macromolecules: Flory's *contact energy* approach and a *cohesive energy density* approach developed independently by Hildebrand (1933) and Scatchard (1931). Both treatments account for van der Waals forces between molecules.

18.2.2.1 Flory's Contact Energy Formulation Treats the Interaction Between Solvent and Polymer Segments

The contact energy approach is a structured model that uses the same ideas that were applied in the lattice approach to the entropy of mixing. As shown in Fig. 18.2 the structural units of the polymer are labeled in sequence. We will consider a unit labeled #2 on each of two polymer chains. Unit #2 is located next to unit #3 and then #4 followed by #5. Unit #2 is a next neighbor to #3 but it is not a next neighbor to #4 or #5. Solvent molecules are labeled #1. As Fig. 18.2 indicates, units labeled #2 can have next-neighbor interactions that are unit-solvent, intermolecular, or intramolecular in nature. Next-neighbor contacts are named short-range interactions and non-next-neighbor interactions are called long-range interactions. The short-range interactions are labeled 1-1, 1-2, 2-2 and the work associated with bringing these units together (using our nomenclature from earlier) will be $w_{\text{s-s}}, w_{\text{s-p}}, w_{\text{p-p}}$. The energy associated with mixing the polymer and solvent molecules is $\Delta w_{\text{s-p}}$ which is written as

$$\Delta w_{\text{s-p}} = w_{\text{s-p}} - \frac{1}{2}\left(w_{\text{s-s}} + w_{\text{p-p}}\right) \tag{18.24}$$

The total number of 1-2 pairs in contact requires defining the number of solvent molecules, n_{s}, the coordination number of the lattice, z, and the number of segments in the polymer chain, y. y will be given by the ratio of the molar volume of polymer to solvent, i.e., $y = \frac{V_{\text{p}}}{V_{\text{s}}}$, viz,

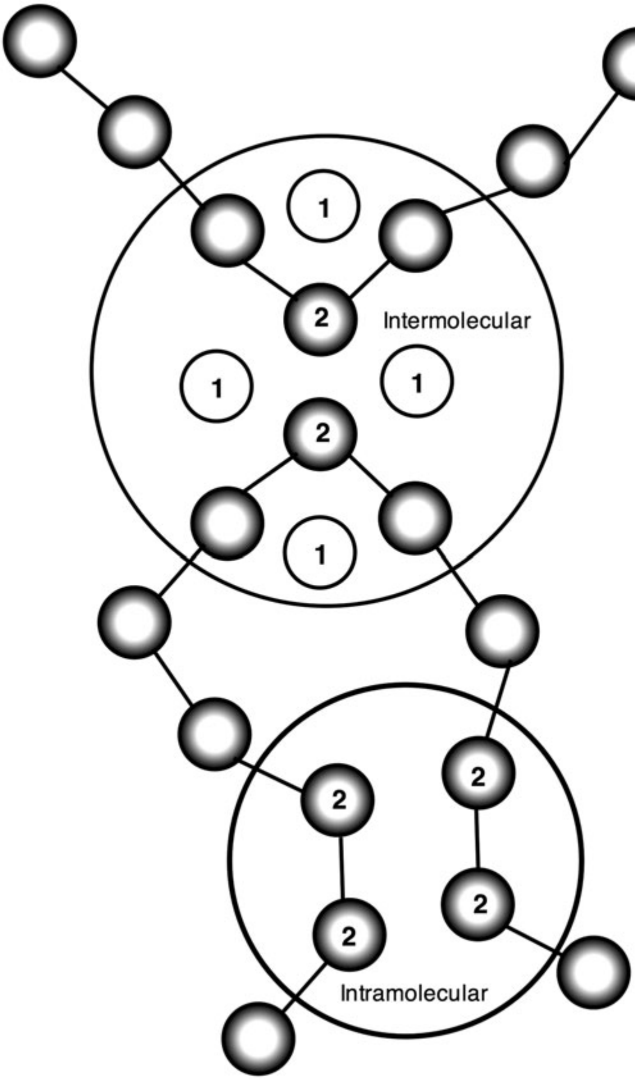


Fig. 18.2 Contact energy model for enthalpy of mixing

$$zyn_{\mathrm{s}}\phi_{\mathrm{p}} \equiv zn_{\mathrm{s}}\phi_{\mathrm{s}} \tag{18.25}$$

The total enthalpy of mixing, ΔH_{mix}, is found by multiplying this number of contacts by $\Delta w_{\mathrm{s-p}}$ with the following result:

$$\Delta H_{\mathrm{mix}} = z\Delta w_{\mathrm{s\,-\,p}}y_{\mathrm{ps}}n_{\mathrm{s}}\phi_{\mathrm{p}} \tag{18.26}$$

Here y_{ps} is the number of polymer segments that are in contact with the solvent. Equation (18.27) is the van Laar expression whose work formed the starting point for both Flory's and Hildegard's formulations. The problem with Eq. (18.27) is that $\Delta w_{\mathrm{s\text{-}p}}$ cannot be calculated. Flory approached this limitation by expressing this interaction energy in terms of a variable, χ_{f} (we will use the subscript f for Flory):

$$\chi_{\mathrm{f}} = \frac{z\Delta w_{\mathrm{s\text{-}p}} y_{\mathrm{ps}}}{kT} \tag{18.27}$$

χ_{f} is a dimensionless variable that characterizes the interaction energy per solvent molecule divided by kT. The physical meaning of χ_{f} is that it is a representation of the energy of the solvent in solution minus the energy of the pure solvent. χ_{f} is experimentally measurable and is useful for investigation of polymer solutions. ΔH_{mix} is related to this more useful observable:

$$\Delta H_{\mathrm{mix}} = kT\chi_{\mathrm{f}} n_{\mathrm{s}} \phi_{\mathrm{p}} \tag{18.28}$$

This enthalpy term will be used shortly in our treatment of the free energy of mixing.

18.2.2.2 The Cohesive Energy Density Parameter Captures the Interactions of Components of a Mixture

While the contact energy approach explicitly considers the lattice structure and in χ_{f} captures the interactions, each χ_{f} must be specifically determined for each solvent–polymer pair under each condition of temperature, polymer molecular weight, and solution composition. A different approach is taken in the cohesive energy density formulation. Here a cohesive parameter term such as the *Hildebrand solubility parameter*, δ, can be used that is a property of just one of the components of the solution. δ is also a term that reflects all of the components of the van der Waals forces and is related to the cohesive energy density, *ced*, of the solvent. *ced* is related to the heat of vaporization. At the temperature of vaporization of a substance, the heat added to release the molecules from condensed phase to vapor phase is a measure of the intermolecular stickiness (cohesiveness). The more easily vaporized the less cohesive the interactions. We can write

$$ced = \left(\frac{\Delta H_{\mathrm{vap}} - RT}{V_{\mathrm{m}}}\right) \tag{18.29}$$

where R, T, and V_{m} are the gas constant, temperature, and molar volume, respectively. *ced* is a numerical value that represents the energy density holding the molecules in a condensed phase together. Therefore, *ced* is a direct measure of the van der Waals interactions between molecules in the liquid phase. The physical intuition that underlies why the heat of vaporization is related to a solubility parameter is that the attractive forces between solvent molecules that must be overcome in the process of vaporization and allowing a solute to dissolve in a solution are the same. When two substances mix, attractive forces between the pure substances must be overcome allowing them to separate, a process that is equivalent whether the molecules go into vapor phase or into solution phase.

How is the enthalpy of vaporization related to the molar internal energy, U, which is the internal energy of a mole of material, j, in condensed phase relative to the same material in ideal vapor form at the same temperature? U is always positive valued and –U is the molar cohesive energy of the condensed phase. –U represents the stabilizing (cohesive) energy in the condensed phase. The internal energy and enthalpy of vaporization are not the same, by definition, $H = (U + pV)$ and for the change between phase 1 (a liquid) and 2 (a perfect gas), both at the same T and P

$$\begin{aligned} &U_1 = H_1 - pV_1 \\ &U_2 = H_2 - pV_2 \\ &\Delta U = U_2 - U_1 = H_2 - H_1 - p\,(V_2 - V_1) \end{aligned} \tag{18.30}$$

The volume difference between a mole of liquid and ideal gas phases is such that $V_2 \gg V_1$ and by the ideal gas law

$$pV_j = RT \tag{18.31}$$

The internal energy change for one mole of j at T can be written as

$$\left(\Delta U_{\text{vap}}\right)_T = \left(\Delta H_{\text{vap}}\right)_T - RT \tag{18.32}$$

The experimental measurement of ΔH_{vap} through Eq. (18.32) allows the calculation of a measure of intermolecular interaction strength in liquid. When this internal cohesive energy is differentiated on the basis volume $(\partial U/\partial V)_T$ is the result which is defined as the *internal pressure* and is closely related to and shares the same dimensions as the cohesive energy density (i.e., MPa or J /cm^3). The internal pressure π_{int} of a substance such as a liquid j is written as

$$\pi_{\text{int}}\,(\text{liq}, j) = (\partial U/\partial Vw)_T \tag{18.33}$$

The internal pressure of a liquid is on the order of 10^8 Pa. If $(\partial U/\partial V)_T$ is rewritten in terms of $\left(\Delta U_{\text{vap}}\right)_T$ and the molar volume of the liquid at the same temperature equation (18.29) is the result. The physical interpretation of both Eqs. (18.33) and (18.29) is of an attractive force energy given by $\left(\Delta U_{\text{vap}}\right)_T$ divided by repulsive interactions which take up space and keep the molecules apart.

The mathematical treatment that connects these physical processes and that leads to the enthalpy of mixing can be appreciated by recalling the first case in which an accounting for the interactions of molecules that forced deviations from the ideal behavior was captured by the van der Waals gas law (Chapter 6):

$$\left(p + \frac{n^2 a}{V^2}\right)(V - nb) = nRT \tag{18.34}$$

This equation specifically accounts for the attractive interactions between molecules and repulsive interactions due to their size. The van der Waals gas constant, a, is a measure of attractive interaction while b is a measure of the excluded volume occupied by a mole of atoms. The dimension of a is Pa/V^2 and that of b is m^3/mol. In Chapter 6, we noted an important limitation of the ideal gas law was that it could not account for a change in phase. The van der Waals equation does treat the condensation of a gas into the condensed liquid phase and conversely the vaporization of the liquid to the gas phase.

In the early twentieth century, van Laar developed an equation to compute the enthalpy of mixing of a binary liquid that had a similar form to the van der Waals equation:

$$\Delta H_{\text{mix}} = \frac{n_1 v_1 n_2 v_2}{n_1 v_1 + n_2 v_2}\left(\frac{a_1^{1/2}}{V_1} - \frac{a_2^{1/2}}{V_2}\right) \tag{18.35}$$

The subscripts in $n_1 : v_1, n_2 : v_2$ refer to the solvent and solute, respectively. It was from this equation that Scatchard and Hildebrand independently derived a treatment of the energy of solvent–solute mixing. This new equation has the same form as the van Laar equation and relates the change in internal energy ΔU to the specific molar volumes and to the energy of vaporization, ΔU^v.

$$\Delta U = (n_1 V_1 + n_2 V_2)\left[\left(\frac{\Delta U_1^v}{V_1}\right)^{1/2} - \left(\frac{\Delta U_2^v}{V_2}\right)^{1/2}\right]^2 \phi_1 \phi_2 \tag{18.36}$$

We recognize in this equation the cohesive energy and the cohesive energy density that have just been discussed and note that the attractive interaction energy term, a, in van der Waals and van Laar equations, is cohesive energy. The Hildebrand parameter, δ, or the solubility parameter can be defined from the following equation:

$$\delta = c^{1/2} = (-U/V)^{1/2} \approx (\Delta U/V)^{1/2} \approx (\Delta H - RT/V)^{1/2} = \frac{a^{1/2}}{V} \tag{18.37}$$

Hildebrand combined Eq. (18.36) with the definition of δ to relate the heat of mixing of a polymer and solvent (or two solvents) to the measurable solubility parameters:

$$\Delta H_{\text{mix}} = V_{\text{m}} \left(\delta_1 - \delta_2\right)^2 \phi_1 \phi_2 \tag{18.38}$$

V_{m} is the volume of the solution. The Hildebrand solubility parameter for materials can be determined following measurement of the refractive index or density. Some of these values are given in Table 18.1. A rough estimate of the miscibility of a solvent and polymer can be made when the following conditions are met:

$$(\delta_2 - 1.1) < \delta_1 < (\delta_2 - 1.1) \tag{18.39}$$

Table 18.1 Hildebrand solubility parameters (δ/MPa$^{1/2}$)

Polymer	δ (SI)
n-Pentane	14.4
Diethyl ether	15.4
Cyclohexane	16.8
Benzene	18.7
Chloroform	18.7
Acetone	19.7
Pyridine	21.7
Dimethylformamide	24.7
Ethanol	26.2
Dimethyl sulfoxide	26.4
Glycerol	36.2
Water	48.0

From Barton (1983) *Handbook of Solubility Parameters*, CRC Press, Boca Raton, FL

The Hildebrand solubility parameter captures in a single value the three components of the van der Waals interaction. Other solubility parameters have been developed to recognize these components to varying degrees. Only one, the Hansen parameter treatment that provides three measures that add up to the Hildebrand parameter, will be mentioned. The Hansen parameters specifically characterize components of

1. dispersion force,
2. hydrogen bonding,
3. polar interactions.

These are related to the Hildebrand δ such that:

$$\delta^2_{\text{Hildebrand}} = \delta^2_{\text{dispersion}} + \delta^2_{\text{H - bonding}} + \delta^2_{\text{polar}} \tag{18.40}$$

The Hansen parameter approach allows each of the attractive components to be charted in a three-dimensional space and with the inclusion of a measure of a radius of interaction create a region of interaction between a given polymer and a particular solvent. Solvents whose parameters lie within the “solubility sphere” will be active solvents for that polymer. Further discussion of these parameters can be found in the references at the end of the chapter.

18.2.3 The Free Energy of Mixing Relates Enthalpy and Entropy in the Standard Manner

We have the necessary state properties defined to directly write the free energy of mixing by combining our analyses of the enthalpy and entropy of mixing, viz., $\Delta G = \Delta H - T\Delta S$.

Using terms of the Flory–Huggins treatment

$$\Delta S_{\text{mix}} = -k\left(n_{\text{s}}\ln\phi_{\text{s}} - n_{\text{p}}\ln\phi_{\text{p}}\right) \tag{18.41}$$

$$\Delta H_{\text{mix}} = kT\chi_{\text{f}}n_{\text{s}}\phi_{\text{p}} \tag{18.42}$$

$$\Delta G_{\text{mix}} = kT\left[\chi_{\text{f}}n_{\text{s}}\phi_{\text{p}} + \left(n_{\text{s}}\ln\phi_{\text{s}} - n_{\text{p}}\ln\phi_{\text{p}}\right)\right] \tag{18.43}$$

In terms of the Hildebrand (and van't Hoff) approach

$$\Delta S_{\text{m}} = -k\left(n_{\text{s}}\ln x_{\text{s}} + n_{\text{p}}\ln x_{\text{p}}\right) \tag{18.44}$$

$$\Delta H_{\text{mix}} = V_{\text{m}}\left(\delta_{\text{s}} - \delta_{\text{p}}\right)^2\phi_{\text{s}}\phi_{\text{p}} \tag{18.45}$$

$$\Delta G_{\text{mix}} = V_{\text{m}}\left(\delta_{\text{s}} - \delta_{\text{p}}\right)^2\phi_{\text{s}}\phi_{\text{p}} + kT\left(n_{\text{s}}\ln x_{\text{s}} + n_{\text{p}}\ln x_{\text{p}}\right) \tag{18.46}$$

18.2.4 Calculation of the Partial Specific Volume and Chemical Potential

In the thermodynamic treatment of solutions the number of moles of each component is needed in addition to *P*, *V*, and *T*. The treatment of macromolecular solutions makes use of chemical potential (u_i), the partial molar volume ($\overline{V_i}$), and partial specific volume ($\overline{v_i}$), in that order:

$$u_i = \left(\frac{\partial G}{\partial n_i}\right)_{T,P,n_i\neq n_j}, \overline{V_i} = \left(\frac{\partial V}{\partial n_i}\right)_{T,P,n_i\neq n_j}, \overline{v_i} = \left(\frac{\partial v}{\partial w_i}\right)_{T,P,w_i\neq w_j} \tag{18.47}$$

The partial molar volume is expressed in volume/mole while partial specific volume is expressed in volume/gram. The partial specific volume can be useful in estimating the volume of a biomacromolecule such as a protein though this is a largely empirical treatment of space occupancy. In the case of a protein, the partial specific volume of a protein, $\overline{v_p}$, can be described in terms of the partial specific volumes of the amino acids, $\overline{v_i}$, and the weight percent, $\overline{w_i}$, of the *i*th residue in that protein.

$$\overline{v_p} = \frac{\sum \overline{v_i}w_i}{\sum w_i} \tag{18.48}$$

Table 18.2 lists the partial specific volume of the amino acid residues as well as several other biopolymers.

Table 18.2 Partial specific volume of amino acid residues

Amino acid	$\overline{v_i}$ (ml/g)	Amino acid	$\overline{v_i}$ (ml/g)
Isoleucine	0.90	Lysine	0.82
Leucine	0.90	Arginine	0.70
Valine	0.86	Glutamic acid	0.66
Alanine	0.74	Aspartic acid	0.60
Threonine	0.70	Methionine	0.75
Serine	0.63	Cysteine	0.61
Phenylalanine	0.77	Proline	0.76
Tryptophan	0.74	Hydroxyproline	0.68
Tyrosine	0.71	Glutamine	0.67
Histidine	0.67		

From Sun S.F. (1994) *Physical Chemistry of Macromolecules*, Wiley, New York

18.2.4.1 The Chemical Potential of Macromolecular Solutions Is Reflected in the Activity of Components

The free energy dependence of a solution on its composition is captured in the chemical potential and measured by the activity or effective concentration. Given our discussion up to this point the interaction between polymers (and polymer segments) and solvent reflected in the partial molar Gibbs free energy, $\overline{\Delta G}_{\text{mix}}$, would be expected to substantially alter the activity compared to a small molecule solution.

$$\Delta u_i = u_i - u_i^{\text{o}} = \overline{\Delta G}_{\text{mix}} = \left(\frac{\partial G_{\text{mix}}}{\partial n_i} \right)_{T,P} \tag{18.49}$$

Remembering that

$$\begin{aligned} u_i &= u_i^{\text{o}} + RT \ln a_i \\ u_i &= u_i^{\text{o}} + RT \ln \gamma_i x_i \end{aligned} \tag{18.50}$$

where i is the ith component, u_i^{o} is the chemical potential in the reference state where $a = 1$and is a function of T and P only, γ_i is the activity coefficient, and x_i is the mole fraction. In a very dilute solution $\gamma_i \to 0$ so Eq. (18.50) can be simplified:

$$u_i = u_i^{\text{o}} + RT \ln x_i \tag{18.51}$$

The two component systems that concern us in solvent–polymer solutions have $x_{\text{s}} \gg x_{\text{p}}$ and thus

$$\ln x_{\text{s}} = \ln \left(1 - x_{\text{p}}\right) = -x_{\text{p}} - \frac{x_{\text{p}}^2}{2} - \frac{x_{\text{p}}^3}{3} - \frac{x_{\text{p}}^4}{4} - \cdots \tag{18.52}$$

This factor, now in terms of the polymer molar fraction, is substituted back into Eq. (18.51):

$$\Delta u_{\mathrm{solvent}} = -RT\left(x_{\mathrm{p}} + \frac{x_{\mathrm{p}}^2}{2} + \frac{x_{\mathrm{p}}^3}{3} + \frac{x_{\mathrm{p}}^4}{4} + \cdots\right) \tag{18.53}$$

The mole fraction of the polymer can be converted to a polymer concentration (in grams/milliliter):

$$x_{\mathrm{p}} = \frac{n_{\mathrm{p}}}{n_{\mathrm{s}} + n_{\mathrm{p}}} = \left(\frac{{}^{\mathrm{g}}\!/_{FW_{\mathrm{p}}}\big/\mathrm{mL}}{(n_{\mathrm{s}}+n_{\mathrm{p}})\big/\mathrm{mL}}\right) \tag{18.54}$$

FW_{p} is the molecular weight of the polymer. Because we started with the assumption that $n_{\mathrm{s}} \gg n_{\mathrm{p}}$, Eq. (18.54) can be simplified such that $n_{\mathrm{s}} + n_{\mathrm{p}} \simeq n_{\mathrm{s}}$, thus:

$$x_{\mathrm{p}} = \frac{{}^{\mathrm{g}}\!/\mathrm{mL}}{FW_{\mathrm{p}}}\frac{\mathrm{mL}}{n_{\mathrm{p}}} = \frac{c_{\mathrm{p}}}{FW_{\mathrm{p}}}V_{\mathrm{s}}^{\mathrm{o}} \tag{18.55}$$

$V_{\mathrm{s}}^{\mathrm{o}}$ is the molar volume of the solvent in milliliters. This can now be applied to the expression for Δu_{s}:

$$\begin{aligned}\Delta u_{\mathrm{solvent}} &= -RTV_{\mathrm{s}}^{\mathrm{o}}\left(\frac{1}{FW_{\mathrm{p}}}c_{\mathrm{p}} + \frac{V_{\mathrm{s}}^{\mathrm{o}}}{2FW_{\mathrm{p}}^2}c_{\mathrm{p}}^2 + \cdots\right)\\ \Delta u_{\mathrm{solvent}} &= -RTV_{\mathrm{s}}^{\mathrm{o}}\left(\frac{1}{FW_{\mathrm{p}}} + A_2c_{\mathrm{p}} + A_3c_{\mathrm{p}}^2 + \cdots\right)\end{aligned} \tag{18.56}$$

This expression has the form of the virial equation for a mole of gas:

$$PV = \left(1 + A_2P + A_3P^2 + A_4P^3 + \cdots\right) \tag{18.57}$$

and therefore the coefficients in Eq. (18.56) can be identified as $A_2 =$ the second virial coefficient and so on. The physical implications of this result can be referenced to the experimentally accessible osmotic pressure and the colligative properties of these solutions.

The final insight that we will examine derives from differentiating the expression derived for the free energy of mixing with respect to the solvent molecules. In terms of the activity of the solvent u_{s},

$$\ln a_{\mathrm{s}} = \frac{\Delta u_{\mathrm{s}}}{kT} \tag{18.58}$$

Now writing in terms of Eq. (18.43):

$$\Delta G_{\text{mix}} = kT\left[\chi_{\text{f}} n_{\text{s}} \phi_{\text{p}} + \left(n_{\text{s}} \ln \phi_{\text{s}} - n_{\text{p}} \ln \phi_{\text{p}}\right)\right] \tag{18.43}$$

$$\ln a_{\text{s}} = \ln\left(1-\phi_{\text{p}}\right) + \left(1-\frac{1}{y}\right)\phi_{\text{p}} + \chi_{\text{f}}\phi_{\text{p}}^2 \tag{18.59}$$

or equivalently following the multiplication by Avogadro's number

$$u_{\text{s}} - u_{\text{s}}^{\text{o}} = RT\left[\ln\left(1-\phi_{\text{p}}\right) + \left(1-\frac{1}{y}\right)\phi_{\text{p}}\right] + RT\chi_{\text{f}}\phi_{\text{p}}^2 \tag{18.60}$$

$\ln\left(1-\phi_{\text{p}}\right)$ can be expanded as before with the result

$$u_{\text{s}} - u_{\text{s}}^{\text{o}} = RT\left[\left(\frac{1}{2}-\chi_{\text{f}}\right)\phi_{\text{p}}^2 + \frac{\phi_{\text{p}}^3}{3} + \cdots\right] \tag{18.61}$$

Flory and Krigbaum approached the physical meaning of this equation by recognizing that the chemical potential is equivalent to the partial molar free energy, so

$$u_{\text{s}} - u_{\text{s}}^{\text{o}} = \Delta\overline{G}_{\text{s}} = \Delta\overline{H}_{\text{s}} - T\Delta\overline{S}_{\text{s}} \tag{18.62}$$

where $\Delta\overline{H}_{\text{s}}$ and $\Delta\overline{S}_{\text{s}}$ are the partial molar enthalpy and entropy, respectively. Two new parameters, a heat of dilution, κ_1, and an entropy of dilution, ψ_1, can be described:

$$\begin{aligned}\Delta\overline{H}_{\text{s}} &= RT\kappa_1\phi_{\text{p}}^2 \\ \Delta\overline{S}_{\text{s}} &= R\psi_1\phi_{\text{p}}^2\end{aligned} \tag{18.63}$$

The relationship in Eq. (18.62) can be rewritten in these new terms to give

$$\begin{aligned}\Delta u_{\text{s}} &= RT\kappa_1\phi_{\text{p}}^2 - RT\psi_1\phi_{\text{p}}^2 \\ \Delta u_{\text{s}} &= -RT\left(\psi_1 - \kappa_1\right)\phi_{\text{p}}^2\end{aligned} \tag{18.64}$$

For a dilute solution, the activity of the solvent is

$$\ln a_{\text{s}} = \left(\kappa_1 - \psi_1\right)\phi_{\text{p}}^2 \tag{18.65}$$

The expression of the difference between the new dilution parameters is seen in Eq. (18.61) and can be equated:

$$\kappa_1 - \psi_1 = \chi_\mathrm{f} - \frac{1}{2} \tag{18.66}$$

Thus Flory and Krigbaum pointed out that the χ_f is composed of two parts, the enthalpy and the entropy. The physical interpretation is that individual polymer chains are isolated and are each surrounded by regions of solvent molecules. The segmental density is no longer considered uniform and the physical structure of the solution is of a dispersion of polymer-segment interacting “clouds” surrounded by regions of pure solvent. This leads to the definition of an ideal temperature θ, at which there is an ideal relationship between the polymer and a particular solvent.

$$\theta = \frac{\kappa_1 T}{\psi_1} \tag{18.67}$$

The activity of the solvent in a macromolecular solution can then be written in these terms:

$$\ln a_\mathrm{s} = -\psi_1 \left(1 - \frac{\theta}{T}\right) \phi_\mathrm{p}^2 \tag{18.68}$$

Thus the activity of the solvent approaches unity (ideality) as the θ temperature is approached.

What picture has emerged of how a polymer interacts with a particular solvent? Solvents are classified as “good” or “poor” and the experimental parameter χ_f helps with the indication. Good solvents have a χ_f value <0.5 and the value of a poor solvent will be >0.5. The θ temperature is a point at which a poor solvent approaches ideal behavior. In fact, good solvent can also have a θ temperature but this temperature is often of high magnitude and therefore not usually measured in the laboratory. What is the physical interpretation of these solvent classifications?

In a *good solvent* the polymer chain will be expanded because the interactions of the polymer with the solvent are energetically favored over the interactions of the polymer with itself. This keeps the polymer in an extended and expanded conformation with space for solvent molecules to approach and pass through “holes” or pass-throughs in the polymer chain. The polymer chain segments compete for space and do not interact, developing instead repulsive interactions. The result of these repulsions is the excluded volume of the polymer, which in a good solvent is substantial and will lead to the maximal extension of the polymer in solution. We will comment on the conformation measures in an upcoming section.

In *poor solvents* the interactions between the polymer units are much preferred compared to interactions with the solvent. The polymer contracts to maximally exclude contact with the solvent. This increases the attraction between polymer segments, and the excluded volume of a polymer in a poor solvent decreases. Polymers in poor solvents exist in a compact form. As the polymer shrinks into its minimal excluded volume it will approach the *theta* (θ) state. For a given homopolymer certain solvents exist which are relatively poor but have a unique quality. At a particular temperature θ, the polymer shrinks to exclude solvent and a balance between the intrapolymer attractive force and the intrapolymer repulsive forces is found. In these solvents the contraction of the polymer is exactly matched by the expanding effect of the excluded volume. The excluded volume effect vanishes. The thermodynamic interpretation for this physical character is that the solvent and solute do not interact, i.e., this is a solution with ideal behavior and can be treated with the formulations of an ideal solution. In fact, the state is pseudo-ideal and is called the *unperturbed state*.

18.2.5 Vapor Pressure Measurements Can Experimentally Be Used to Indicate Interaction Energies

Now we have both thermodynamic expressions and molecular interpretations for the entropy, enthalpy, free energy, and activity of polymer solutions. It is straightforward to write expressions for experimentally accessible observables such as the colligative properties and vapor pressure of these solutions. Vapor pressure measurements made over a range of polymer concentrations are an effective experimental method for obtaining ΔG, ΔH, and ΔS. If the behavior of a solvent is characterized by its activity over a concentration range (where P_1° is the vapor pressure of the pure solvent and P_1 the pressure over the appropriate solution), then

$$\ln a_1 = \frac{P_1}{P_1^\circ} = \frac{\mu_1 - \mu_1^\circ}{R'T} \tag{18.69}$$

which after Flory's treatment can be rewritten as

$$\ln a_1 = \ln(1 - \phi_2) + \left(1 - \frac{1}{y}\right)\phi_2 + \chi_1\phi_2^2 \tag{18.70}$$

ϕ_1 and ϕ_2 are known from the experimental preparation and y can be calculated from V_1 and V_2. χ_1 can be calculated once a_1 is determined experimentally. $\Delta\overline{H_1}$ and $\Delta\overline{S_1}$ can be calculated from χ_1 or from the temperature coefficient of the activity a_1.

$$\begin{aligned}\Delta\overline{H_1} &= R'T\chi_1\phi_2^2 = -R'T\left(\frac{\partial \ln a_1}{\partial T}\right)_{P_1,\phi_2} \\ \Delta\overline{S_1} &= -R'\left[\ln(1-\phi_2) + \left(1 - \frac{1}{y}\right)\phi_2\right] = -R\left(\frac{\partial\,(T \ln a_1)}{\partial T}\right)\end{aligned} \tag{18.71}$$

These experimentally determined values of χ_1 provide a measure of the interaction between solute and solvent in the dilute polymer solution.

18.3 The Conformation of Simple Polymers Can Be Modeled by a Random Walk and a Markov Process

The simplest model of a polymer's chain configuration is to consider a chain composed of equal length segments that at each junction point are freely movable in all directions. This idealized model treats each segment as a line without thickness, and so the excluded volume (the space taken up by one segment that cannot be occupied by a second segment) is ignored. To find the possible chain configurations a random walk process is used. The method of the random walk will be detailed in our discussion of diffusion in Chapter 21. The random walk is similar to a set of Bernoulli trials. The sequence of steps is characterized by the independence of each trial from the others. Markov explored the general case of a random or stochastic process in which the outcome of any trial depends only on the current state. Such a theory appropriately describes a polymer chain and is commonly referred to as "random flight." The approach is to consider a sequence of monomers that are represented by a set of vectors $\boldsymbol{I_1}$, $\boldsymbol{I_2}$,. . . that are joined at atoms or molecules that are represented by a set of points $\boldsymbol{M_1}$, $\boldsymbol{M_2}$,. . .. The *i*th bond connects the (*i* –1)th and *i*th atoms in the chain and is characterized by vector $\boldsymbol{I_\iota}$ which has magnitude a_i. The polymer made from (n +1) monomers is therefore characterized by a set of n vectors $\boldsymbol{I_\iota}$ at any given moment. There are no constraints or inter-relationships between the vectors (monomers) as currently written so the description will be only a general geometrical one. The distance between ends of the polymer chain can be calculated from the vector $\boldsymbol{r}$ that connects the two ends of the chain:

$$r = \sum_i Ii \tag{18.72}$$

The scalar mean value r^2 is required for a strictly statistical description:

$$r^2 = \sum_{i,j} Ii \cdot Ij \tag{18.73}$$

The end-to-end distance of the chain will be proportional to the segment length and the square root of the number of segments, $n^{1/2}$. There will be a large ensemble of different conformations, all of the same energy. The statistical behavior of a random coil is consistent with a Gaussian distribution of the segments around the center of mass. The measure of the average distance from the center of mass to any given unit is called the *radius of gyration*, $\mathbf{R_G}$, and gives a sense of dimension to the volume occupied by the macromolecule. Thus for a true random flight or chain, R_G will have the $n^{1/2}$ relationship.

A natural model of polypeptide or nucleic acid chains includes a factor for chain "stiffness" reflecting the bond angles and finite atomic dimensions of the segments and their connections. This model is described as a worm-like chain and takes into account the excluded volume of the chain. P. Flory showed that the volume occupied by a chain of finite size will increase with chain length as $n^{2/3}$ rather than by $n^{1/2}$.

The apparent volume of a chain of a particular length depends on the chemical nature of the polymer and its interactions with the solvent as has been discussed earlier.

What is the size relationship of the polymer in different types of solvents? In good solvents, the excluded volume effect will be maximal and the polymer chain will be in an extended form greater than $n^{1/2}$. In poor solvents, the polymer moves toward a minimal excluded volume and depending on the temperature may or may not be in the θ state. In the θ state, the polymer will have the $n^{1/2}$ size relationship of a Markov chain. For a heteropolymer like a polypeptide no single solvent will be good, poor, or θ. Thus the solvent will have a variable effect on the polymer conformation altering the compactness and stiffness of the chain locally. H_2O is generally considered a poor solvent in relationship to the stable native configuration of most proteins. This is apparent since there is relatively little water found in the interior of most proteins.

Solutions of biological macromolecules can usually be treated as having one of only several geometries, the random coil, rod-like, or globular shape. Molecules that have a rod-like conformation in solution usually have a helical secondary structure and include the α-helices of polypeptides and the helical structures of the polynucleotides. The properties of these solutions are dominated by the length of the rod rather than by the diameter. The compact conformation found in poor solvents results from a positive entropic term due to the exclusion of solvent and a negative enthalpic term due to the attractive side-chain interaction. These conformations give an overall globular quality to the macromolecule, and the shape can be represented by a sphere or ellipsoid of prolate (cigar shaped) or oblate (flattened sphere) shape.

18.4 The Major Classes of Biochemical Species Form Macromolecular Structures

Most of the molecular constituents of biological systems including carbohydrates, amino acids, nucleic acids, and lipids form macromolecular structures. Water forms macroarrays as do the ions of calcium and phosphate when forming the extended calcium–phosphorus crystalline structure of bone and teeth. Yet of this list, only the first three polymerize to form macromolecular solutions. We will now briefly review certain aspects of the macromolecular polymers formed from these monomers.

18.4.1 Nucleic Acids Are the Basis for Genetic Information Storage and Processing

Biological systems are complicated, and each element must be synthesized and coordinated before life can be successfully lived. If the organism is to pass on its hard-won evolutionary gains, an efficient and accurate mechanism for passing on the control program must exist. It is in the nucleic acids, the purine and pyrimidine bases of DNA, that this program is stored and used in life and then passed on in reproduction. The instruments that a cell uses to perform its life's work are essentially proteinaceous. It is the proteins that modify and act: on lipids, on carbohydrates, and on the small inorganic molecules to play the symphony that is life. The genes that comprise DNA do not directly cause the production of proteins but act through the closely related molecules, RNA. The code for protein production is written in the DNA; it is transcribed into a messenger RNA (mRNA); and then the mRNA is read by the ribosome. The ribosome is composed of ribosomal RNA (rRNA), and it catalyses the assembly of the protein from constituent amino acids brought to the ribosome by a transfer complex also made from RNA (tRNA). We will only briefly describe the chemical elements that comprise DNA and RNA and point out several of their shared biophysical properties. Starting points for discovering more details are given in the reference section of this chapter.

DNA (*deoxyribonucleic acid*) and RNA (*ribonucleic acid*) are both *nucleic acids*. This means that they are polymers of deoxyribo*nucleotide* and ribo*nucleotide* units, respectively. *Nucleotides* are composed of a *nucleoside* plus one or more phosphate groups which serve as the links in the polymer structure. A nucleoside is composed of a nitrogenous base (either a purine or a pyrimidine) and a pentose sugar: deoxyribose in the case of DNA and ribose in the case of RNA. There are essentially five bases, two sugars and one phosphate from which to form a nucleotide. The ones in DNA are illustrated in Fig. 18.3. The nucleotides are phosphoric acid esters of the nucleosides. The nucleic acids are polymers formed by the polycondensation of mononucleotides bound covalently via phosphodiester bonds between the 3′ hydroxyl group of one nucleotide and the 5′ hydroxyl of the second. The

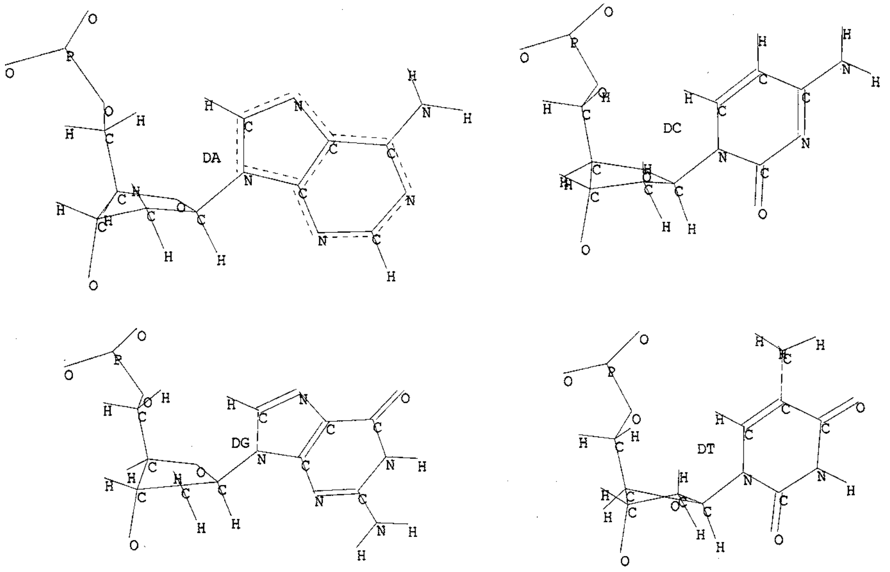


Fig. 18.3 The nucleotides that make up DNA. *Clockwise*: dAMP, dCMP, dGMP, dTMP

backbone of the nucleic acids is thus built from repeating sugar-phosphate-sugar chains. These chains have an acidic nature secondary to the universally negatively phosphate groups which are usually neutralized by Mg^{2+} ions. The nucleic acids are water soluble while the free nitrogenous bases are very poorly soluble in water. There is a linear ordering to the bases on a given chain, and the genetic code is carried by triplets of base order called *codons* (Table 18.3). The primary structure or sequence of the DNA as represented by the codons provides the information for the primary sequence of proteins. The genetic code necessary to provide the primary sequence for a specific protein is called a *gene*. Not all of the DNA molecule codes for proteins. The portions of the DNA base sequences that are transcribed into RNA and ultimately into an amino acid sequence are called *exons*. The untranscribed portions are called *introns*.

Table 18.3 The genetic code

AUU	Ile	GUU	Val	UUU	Phe	CUU	Leu
AUC	Ile	GUC	Val	UUC	Phe	CUC	Leu
AUA	Ile	GUA	Val	UUA	Leu	CUA	Leu
AUG	Met (Start)	GUG	Val	UUG	Leu	CUG	Leu
ACU	Thr	GCU	Ala	UCU	Ser	CCU	Pro
ACC	Thr	GCC	Ala	UCC	Ser	CCC	Pro
ACA	Thr	GCA	Ala	UCA	Ser	CCA	Pro
ACG	Thr	GCG	Ala	UCG	Ser	CCG	Pro
AAU	Asn	GAU	Asp	UAU	Tyr	CAU	His
AAC	Asn	GAC	Asp	UAC	Tyr	CAC	His
AAA	Lys	GAA	Glu	UAA	Stop	CAA	Gln
AAG	Lys	GAG	Glu	UAG	Stop	CAG	Gln
AGU	Ser	GGU	Gly	UGU	Cys	CGU	Arg
AGC	Ser	GGC	Gly	UGC	Cys	CGC	Arg
AGA	Arg	GGA	Gly	UGA	Stop	CGA	Arg
AGG	Arg	GGG	Gly	UGG	Trp	CGG	Arg

The overall structure of DNA derives from the arrangement of the polycondensed nucleotides and the geometric constraints of the components of the nucleotides. A very important property of the purine and pyrimidine bases derives from their ability to form resonance structures based on extensive keto–enol tautomerization. The tautomeric forms are very sensitive to pH due to the multiple hydrogen acceptor sites on the bases. The resonance structures give these bases their UV absorption maxima in the 260–280 nm range with significant variation that is dependent on the pH. The UV absorption of the nucleotides and nucleic acids is due to the nitrogenous bases and varies with secondary structure as well as with pH. The pyrimidines are planar molecules, and the purines are nearly so. The purines and pyrimidines easily form in definite pairs stabilized by hydrogen bonds (Fig. 18.4). These pairs are crucial in the structure and the fundamental activity of the nucleic acids: the preservation and flow of information. In DNA, adenine and thymine form a base pair stabilized by two hydrogen bonds, and guanine and cytosine form a second base pair stabilized by three hydrogen bonds. The formation of these pairs usually occurs between two

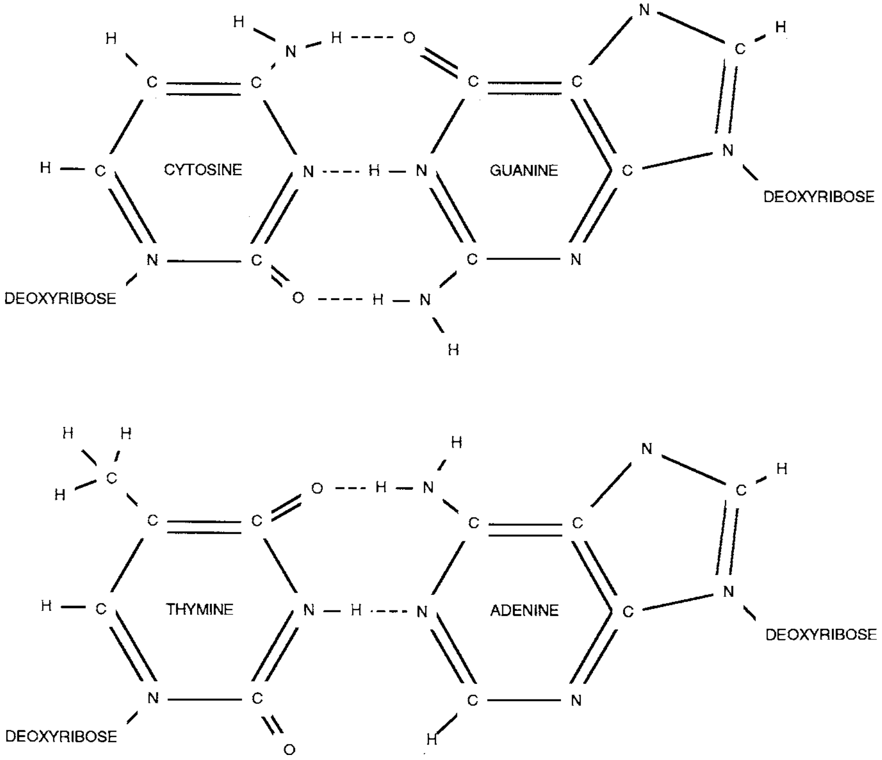


Fig. 18.4 The pyrimidines and the purines form in pairs stabilized by hydrogen bonds

different strands of DNA thus forming two complementary strands that are antiparallel. Because of the constraints of the hydrogen bonding, every purine in one strand corresponds to a pyrimidine in the complementary strand (A + T = G + C; A = T and C = G). These rules of complementarity were established by Chargaff and lead to the idea that DNA is double stranded. If the base pairs are opened by disrupting the hydrogen bonds, the two strands can act as templates either for the production of a complementary RNA transcript (the mRNA) or for a new pair of complementary DNA daughter strands, thus leading to the replication of the DNA molecule. The correct matching of complementary nucleotides to the strand depends on the proper hydrogen bond formation which in turn is dependent on the correct keto–enol structure of the bases. In contrast to DNA, RNA is generally single stranded with most base pairing occurring within the chain itself. Another important difference is the replacement of uracil for thymine in RNA.

The two antiparallel complementary strands of DNA form the basis for the secondary structure of DNA. This structure is the familiar double helix formation described by Watson and Crick. The two strands are wound together around a helical axis giving rise to a double screw-like structure. Figure 18.5 shows the wide

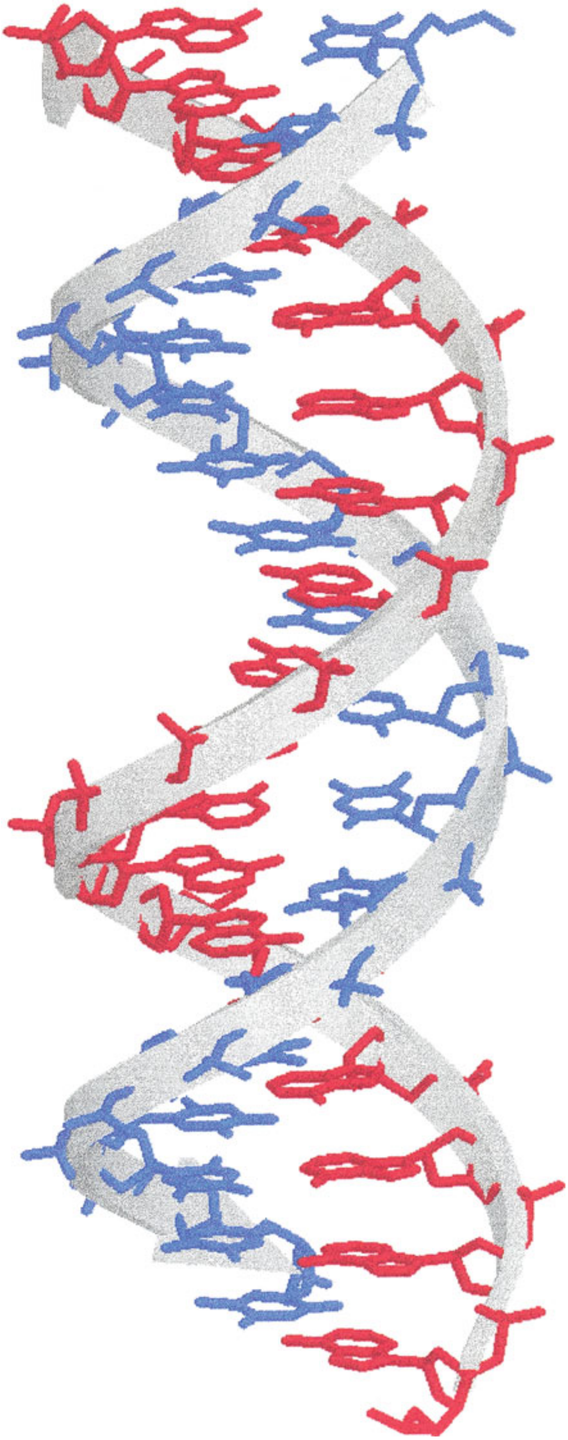


Fig. 18.5 Structure of the double helical DNA. The wide and narrow grooves can be seen in this double twist right-handed helix

and narrow grooves in this double twist right-handed helix. The narrow groove is formed by the side by side chains of the sugar–phosphate backbone. The negatively charged phosphate groups are easily accessible to being neutralized by cations. The base pairs lie within the helix and form the wide groove. The plane of the base pairs lies perpendicular to the helix axis and the vertical distance between bases is 0.34 nm. The nucleotides turn at 36° and there are 10 nucleotides per turn of the helix. A full turn measures 3.4 nm. The deoxyribose sugar moiety is crucial for the formation of the double helix since the additional oxygen found in the ribose moiety makes it sterically impossible for the double helix to form. The bases each have a dipole moment, and so the interaction between base pairs includes dipole–dipole interactions and dispersion or London forces as well as the hydrogen bonds. The net energetics of the G:C pair is significantly attractive because the dipole interactions are attractive and this is added to the hydrogen bonding interaction. The A:T pair repel one another because of the opposing dipole vectors which overcome the energy of the two hydrogen bonds (Table 18.4). These forces therefore make regions containing A:T bases less stable and more likely to be denatured than G:C regions. The stacking of the base pairs leads to interactional energy with pairs above and below a given set of bases (Table 18.5). These stabilizing interactions are due to van der Waals interactions and stabilize the geometry of the bases.

Table 18.4 Base interactional energies (kJ/mol)

Base pair	Hydrogen bonds	London	Total energy
A:T	–26	+1.00	–25
G:C	−40	−16.3	−56.3

There are two hydrogen bonds formed between A-T and A-U pairs while three hydrogen bonds are formed between G-C pairs. The dipole–dipole interactions for A-T and A-U pairs are repulsive while those between the G-C pair are attractive. The total energy of the pair is given

Consideration of these interactional energies helps account for the highly cooperative melting behavior of the nucleic acids which starts at the AT regions. A variety of different DNA secondary structures, the A, B, and Z forms, can be found that take form depending on the hydration state and ionic strength of the environment. These structures are illustrated in Fig. 18.6. The A state occurs when B-DNA is dehydrated and is squatter, with the minor curve almost eliminated. There is a decreased binding of water molecules by the phosphate groups in the A form which is why dehydration drives the conformational form B?A. The Z-form is distinct from both the A and B forms because it is left handed. It only forms in high ionic strength solvents because the electrostatic interaction between the phosphate groups must be neutralized in order to favor the conformation.

Currently a large body of biophysical research is exploring the interactions between DNA, RNA, and proteins. Both the major and minor grooves are lined with potential hydrogen bonding sites. These sites represent patterns of bonding

Table 18.5 Stacking interactional energies

Stacked pairs	Energy (kJ/2 mol of bases)
CG GC	−82.8
GC GC	−49.8
TA CG	−46.9
AT CG	−32.6
AT GC	−31.0
TA GC	−30.1
GC CG	−23.8
TA AT	−21.8
AT AT	−20.9
AT TA	−6.7

The dipole–dipole interactions also lead to interactions between stacked base pairs. These energies are given and need to be added to the paired interactional energies to arrive at the total interaction energy for a set of bases in three dimensions

that seem to provide specific potential energy surfaces that can be "recognized" and interacted with by proteins. In general proteins fit into and read the major groove more easily. For example, the bacterial restriction endonucleases like EcoRV which maintain the integrity of bacterial DNA by searching for certain palindromic hexanucleotides in the foreign DNA recognize the nucleic acid sequence GATATC. The EcoRV protein diffuses along the major groove and a loop of peptide that acts to read GATATC contacts the major groove. If the nucleotide sequence is encountered, the recognition sequence forms six hydrogen bonds with the GAT portion of the nucleotide sequence. Since GATATC is a palindrome, the other DNA strand also has a GAT sequence. EcoRV is in fact symmetrical around the axis of the DNA and reads both strands simultaneously. Thus when it encounters GAT it does so with twofold symmetry and a total of 12 hydrogen bonds are formed. The restriction endonuclease alters shape on binding and kinks the DNA, then binds catalytic Mg^{2+} and cleaves the foreign DNA. How does the EcoRV differentiate between bacterial and viral DNA? The bacterial DNA methylates its own restriction enzyme recognition sites and so its own restriction enzyme will not be able to form hydrogen bonds and fails to recognize the nucleotide sequence, thus appearing to have self-recognition.

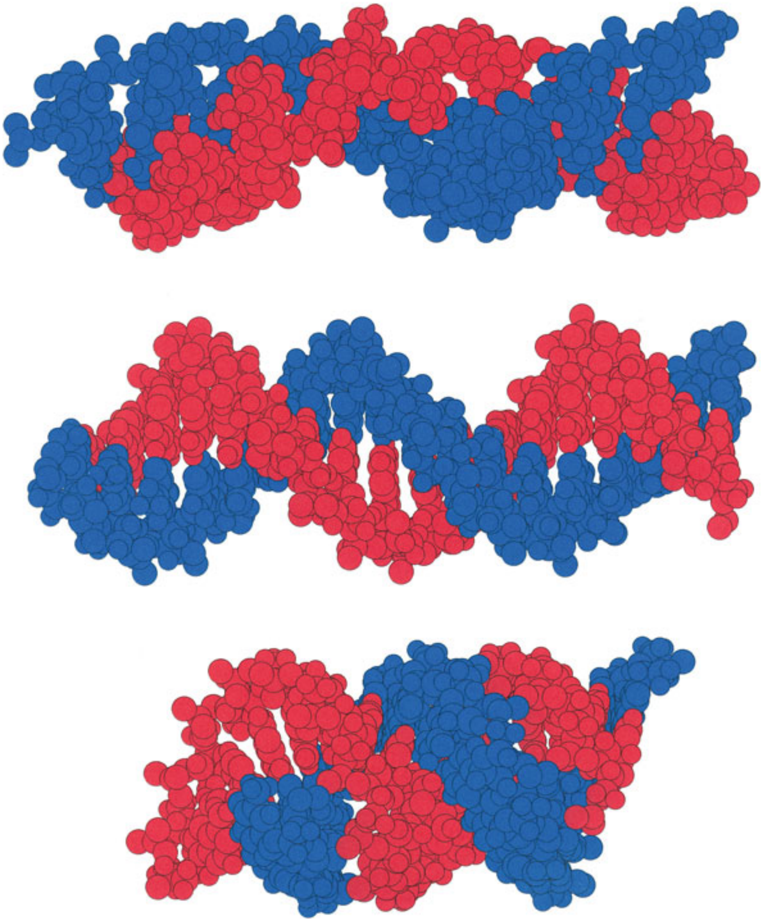


Fig. 18.6 DNA secondary structures, space filling models of A, B, and Z DNA. These forms may be found depending on the hydration state and ionic strength of the environment

18.4.2 Carbohydrate Polymers Are Dominated by Hydrophilic Interactions with Water

Carbohydrates are aldehydes or ketones that are also poly-alcohols. The major saccharides of biological importance are the 3, 5, and 6 aldoses and ketoses. The chemical structure of several of the six-carbon aldose D-isomers is shown in Fig. 18.7. In solution the pentoses and hexoses spontaneously cyclize and form furanose (five-carbon) and pyranose (six-carbon) rings. These rings are formed when the aldehyde or ketone is able to react with an alcoholic group, thus forming a hemiacetal or hemiketal. These ring structures can form a variety of structures, the two

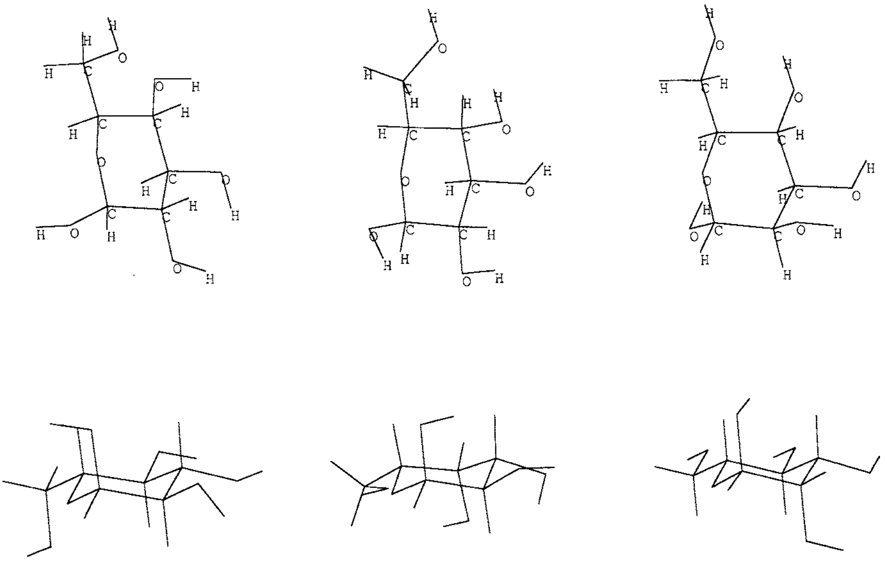


Fig. 18.7 D-isomers of the major aldoses of biological importance: glucose, galactose, and mannose. Note the chair configurations of the cyclic sugars. Steric constraints of the side chains determine the conformation of the pyranose ring

extremes being a "boat" form and a "chair" form for the pyranose ring and a puckered "envelope" form for the furanose rings. Generally the pyranose rings will prefer the chair configuration because there is less steric hindrance than when the bulkier side groups are located equatorially. Some sugars have a negatively charged sulfate or carboxylate group substituted for the hydroxyl group. Positively charged substituents are also found. Carbohydrates of this class are designated as *substituted*. The addition of an ionic group on the sugar allows significant electrostatic interaction to be provided by these sugars. We will see how these complex sugars can significantly affect the mechanical properties of biomaterials as we discuss *polysaccharide* polymers.

Polymers of sugars are usually quite large, with molecular weights ranging from 10^4 to 10^7 Da. These polymers may be either random or highly ordered, and the large polymers of glucose are a good example of how important the organization of the polymer is to the properties of the molecule. The glycosidic linkages between sugar molecules are formed between the hemiacetal group of one monosaccharide and an alcohol group of a second sugar and thus are actually *O*-glycosidic linkages. This reaction forms an acetal with two possible bond orientations. If the bond is formed between the C-1 carbon and the C-4 hydroxyl oxygen such that the bond from the C-1 carbon is above the plane of the ring, the linkage is called a $\beta(1\rightarrow4)$ *glycosidic linkage*. Alternatively, if the linkage is formed by a bond from beneath the plane of the ring, an $\alpha(1\rightarrow4)$ *linkage* will be formed. Glycosidic linkages can be formed between the hemiacetal carbon and the nitrogen of an amine. These linkages are called *N-glycosidic* linkages and play important roles in the formation of the nucleotides that make up DNA and RNA. N-glycosidic linkages are usually β configurations.

The polymeric structures of the three major polymers of glucose form starches, glycogen, and cellulose. Both starch and glycogen are semi-randomly ordered polymers used for energy storage, starch by plants and glycogen by animals. Starch is a mixture of amylose and amylopectin. Amylose is an unbranched polymer of glucose molecules connected in α(1→4) linkages. Amylopectin is α(1→4)-linked glucose molecule with branches formed by an α(1→6) linkage every 30 or so units. Glycogen is similar to amylopectin except that the α(1→6) branches occur more frequently, about every 10 glucose units. Finally cellulose is an unbranched polymer formed entirely from β(1→4) linkages. The orientation of the glucose molecules is rotated 180° so that the ring oxygen hydrogen bonds to the 3-OH group of the neighboring glucose. Long parallel chains of glucose molecules are formed that have enormous tensile strength, thus giving cellulose the properties that make it an important structural polymer in plants. This is in contrast to the hollow helical shape of starch and glycogen. The structure of these energy storage polymers allows easy access to the energy store but provides little structural strength.

Materials with many different properties can be constructed from carbohydrate polymers. An important quality of these polymers is the ability to swell tremendously when exposed to water. The hydroxyl groups and the ionic groups in the polymers of substituted sugars become highly hydrated and thus can act as a biological hydraulic system. For example, in proteoglycans which form the extracellular matrix of cartilage, a large number of chains of glycosaminoglycans are attached to a core protein and are heavily hydrated with the water molecules held tightly by the carbohydrate chains. When this material is deformed, the combination of ionic repulsion and hydration forces (entropic, dipole interactions, and hydrogen bonding) cushion the deforming blow by attempting to quickly restore the interactional energies to a minimum.

18.4.3 The Polymers of Amino Acids, Proteins Are by Far the Most Diverse and Complex of All Biological Polymer Families

Proteins are biological heteropolymers comprised of 20 naturally occurring L-α-amino acids that are joined together by amide bond linkages. The amino acids are organic zwitterions that consist of a central α-carbon *sp*^{3} hybridized to an amino group, a carboxylic acid group, a hydrogen, and a side chain, which gives each amino acid a more or less unique character. The side chains are important in the interactional energies that give amino acids and proteins their far-ranging properties. Much of our attention is directed toward the unique class properties conferred on peptides and proteins by the side chains. Two of the natural amino acids, glycine and proline, stand out as different from the others in the group. Glycine contains two hydrogen atoms bonded to the α-carbon and so has no optical asymmetry. All the other amino acids isolated from proteins are L stereo-isomers. Proline, an α-imino acid, has a side chain that is bonded to both the α-carbon and the amino group. This side chain forms a cyclic ring that induces significant constraints on the backbone structure otherwise found when all of the other amino acids are joined in peptide linkages.

Amino acids are linked to one another by a condensation reaction between the amino and carboxylic acid groups of two amino acids. This condensation reaction leads to the formation of an amide linkage which is a bond with limited rotation around the O=C–N bond. Chains of amino acids linked together by amide bonds form polypeptides. The backbone structure of a polypeptide is an extended series of amide linkages alternating with the sp^3-bonded α-carbon. The steric constraints of a polypeptide chain are severe, as originally demonstrated by Pauling. These constraints arise from

(1) the tetrahedral nature of the carbon atom, which restricts its bonds to specific directions and angles (Fig. 18.8a),

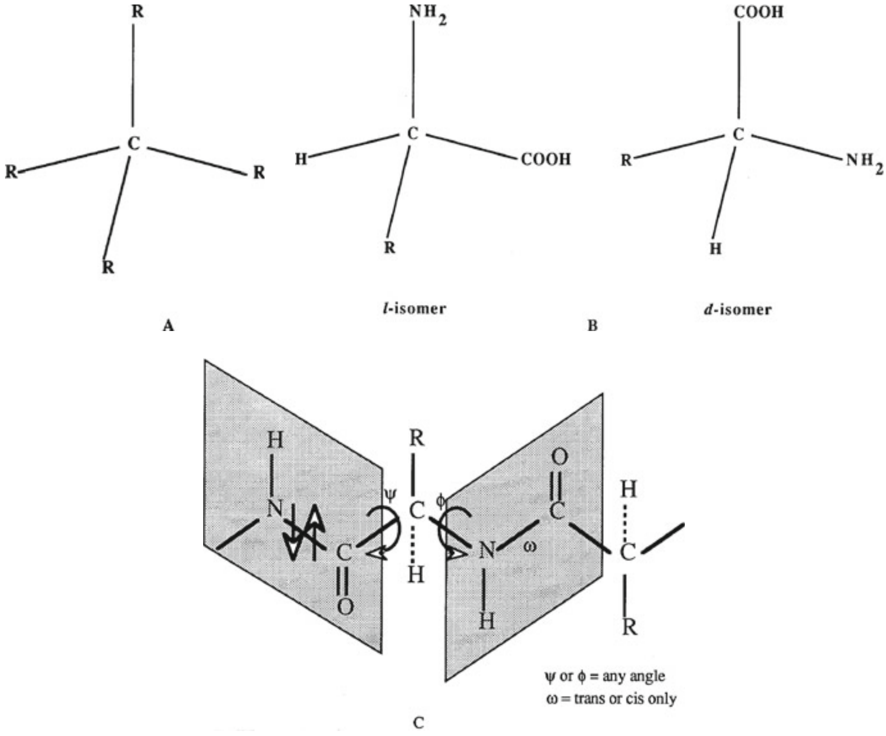


Fig. 18.8 The character of the peptide chain depends on the restrictions imposed by the peptide bond: (**a**) tetrahedral bonding pattern of carbon, (**b**) asymmetry of the α carbon, and (**c**) planar character of the peptide bond

(2) the asymmetry of the α-carbon atom of all amino acids except glycine (Fig. 18.8b), and
(3) the existence of hybrid orbitals in the peptide bond attributable to keto–enol isomerization, which imparts a double bond character restricting rotation about that bond and making the peptide bond a planar entity (Fig. 18.8c).

The orbital structure of the amide bonds creates two binding planes centered around the α-carbons that are free to rotate with respect to one another only at the single bonded N-α-C and C-α-C bonds. The dihedral angle associated with the rotation around N-α-C is labeled ϕ and that associated with the C-α-C is labeled ψ. The geometry of the amino acid-derived peptidyl backbone is shown in Fig. 18.8. There are a limited number of energetically likely angles in which ϕ and ψ will be found due to the steric repulsion of the carbonyl oxygen and the proton bonded to the nitrogen atom. Under these restrictions, only three structures, or some mixture of these conformations, are available to a poly-α-L-amino acid. These three structures are the *helix*, the *pleated sheet*, and the *random chain*. These structures are the stable ones calculated on the basis of the polypeptide "backbone" chain, and their stabilization depends primarily on the intrachain formation of hydrogen bonds. These structures are examples of *secondary* structure. Which secondary structure is energetically preferred depends to a great degree on the *primary* structure, or the order of amino acids in the peptide chain. Depending on the side chains of the amino acids, a variety of interactional energies are called into play, and the local secondary structure is formed from the possible structures. The local secondary structure will be stabilized or destabilized by the interaction of the side chains with each other, the solvent, or other protein, lipid, nucleic acid, or carbohydrate (macro)molecules. It is these "environmental" factors that strongly influence the secondary structure as well as the interaction of these local domains to form the overall folding structure of the protein, the *tertiary* structure. A final level of protein structure can be defined, the *quaternary* structure which describes the association of protein subunits into protein complexes such as hemoglobin or the pyruvate decarboxylate complex.

Both the strongly intrachain hydrogen-bonded helix (Fig. 18.9a) and the extended pleated sheet (Fig. 18.9b) have a high degree of periodicity. The third structure exhibits no periodicity and is called a *random chain*. The more extended the polymer, the greater is the number of H_2O molecules whose state is perturbed. The thermodynamic result of this perturbation depends upon the relative extent of interaction of the polymer with itself versus that with the aqueous solvent. If there is no steric hindrance, the polypeptide will fold so as to maximize the intrapolymer interactions and to expose the least surface to its aqueous surroundings. We will explore these secondary structures in greater detail shortly.

The structure, properties, and functions of proteins depend on the interactional and chemical behavior of the side chain groups. Many of the non-covalent interactions of the side chains can be understood by using the following side chain classification:

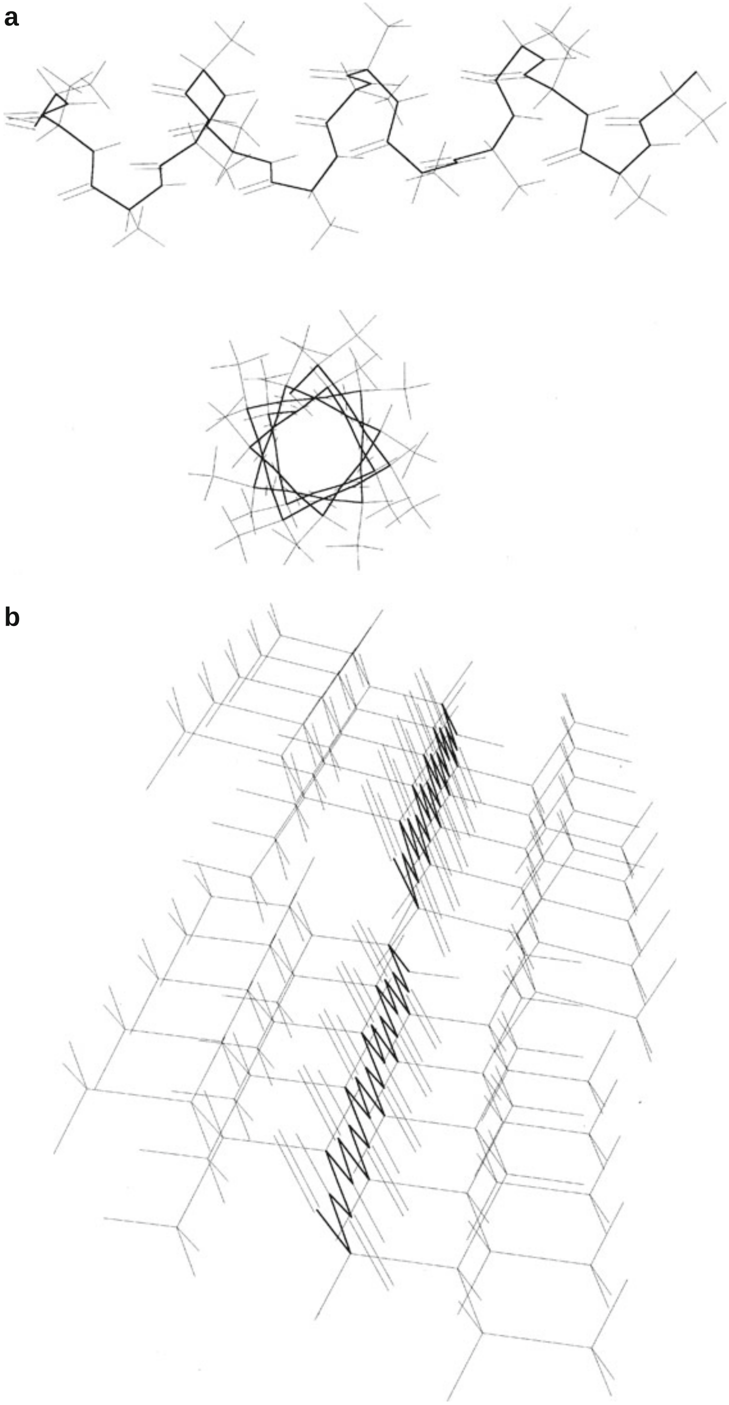


Fig. 18.9

(1) hydrophobic or nonpolar,
(2) neutral (uncharged),
(3) acidic (anionic), and
(4) basic (cationic).

Figure 18.10 provides the chemical structures of the principal amino acids in proteins. The likely role of a given interactional energy can usually be deduced from inspection of the side chain, and even within a grouping we would expect to find variation. For example, the hydrophobicity of amino acids can be ascertained by measuring the ΔG for the transfer from water to organic solvents as shown in Table 18.6. These numbers show that Gly<Ala<Val<Ile which would be the predicted order.

Several of the amino acid side chains contain amino and carboxylic acid moieties. This is in addition to the universal acid–base properties of individual amino acids that are the result of the amino and carboxylic moieties attached to the ?-carbon. In aqueous solution at physiological pH, these groups are ionized, and free amino acids are dipolar zwitterions with a net charge of 0. This is because the pK_a of the carboxylic acid group is 3.5 and the pK_a of the amino group is 10. With a physiological pH around 7 the ionized form of both of these moieties is dominant at a ratio of nearly 1000:1. The titration curves for dipolar amino acids illustrate this clearly (Section 14.6.7). The amino acids of the acidic or basic groups have a net charge of +1 or –1 again because of the pK_as of these groups. In general, only the side chains contribute to the overall net charge of the polypeptide since only the N-terminal and C-terminal amino acid backbone groups are available to be ionized, all other groups having been used in the peptide bond formation. (However, as studies of hemoglobin have shown, the acid–base status of the terminal amino acids can be very important in the biological function of the protein. In the case of hemoglobin a substantial quantity of CO_2 is transported out of the tissues following the carbamation of the terminal amino group of each subunit of the hemoglobin molecule.) The differences in the net charge of a polypeptide at different conditions of pH provide an important method for the sensitive separation of macromolecules through the electrophoretic technique of isoelectric focusing.

Inspection of Table 18.7 reveals that the acid–base behavior of histidine in the range of physiological pH is unique among the amino acids. Because histidine has a pK_a of 6.0, small changes in the local pH in physiological systems can dramatically

Fig. 18.9 (continued) Periodic structures of the peptidyl backbone. (**a**) The alpha helix forms so that the peptidyl backbone spirals around a central line with the side chains on the outside of the structure and an inner core formed by the peptidyl chain. These features can be appreciated in the framework and ball and stick models showing a lengthwise and an end-on view. The backbone is highlighted with the bolder grays. (**b**) In the β-sheet, the backbone is extended with the side chains found alternatively above and below the backbone. In these images the backbones of two parallel chains can be appreciated as they move away from the viewer. The side chains form planes above and below the inner sheet-like laminar peptidyl backbones

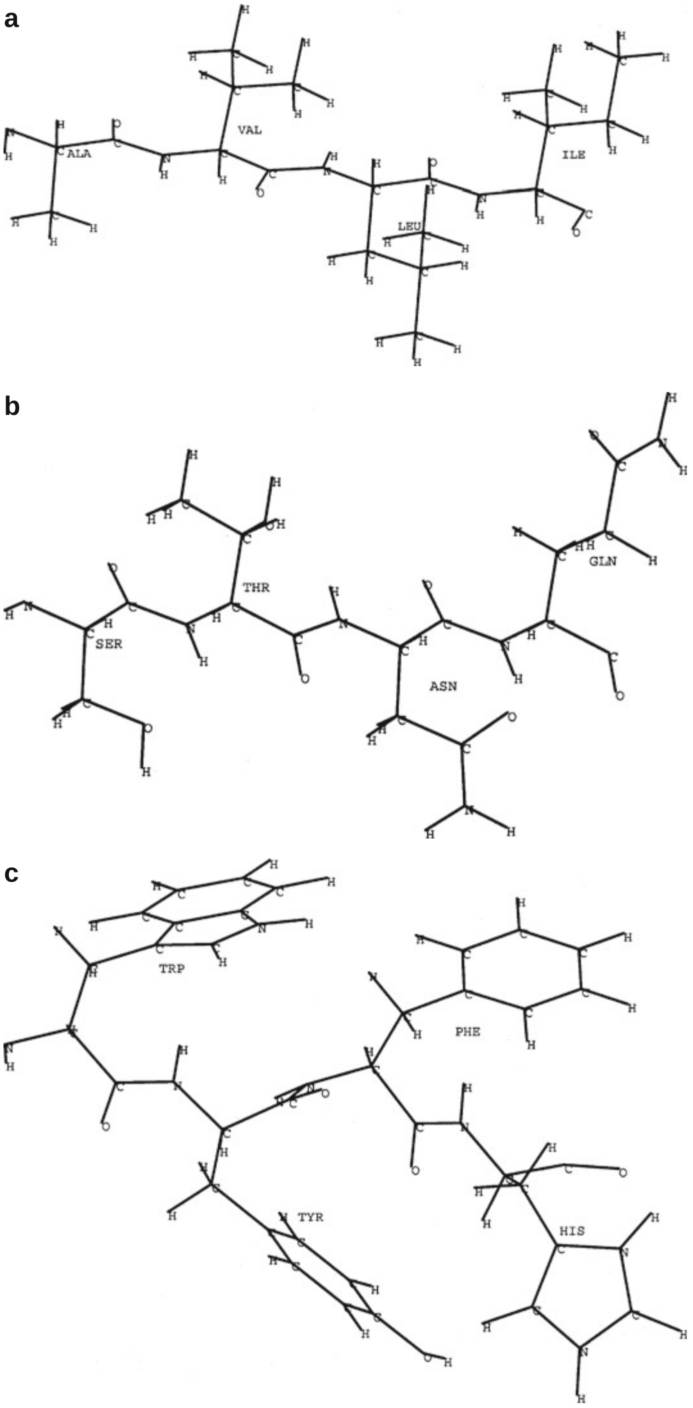


Fig. 18.10

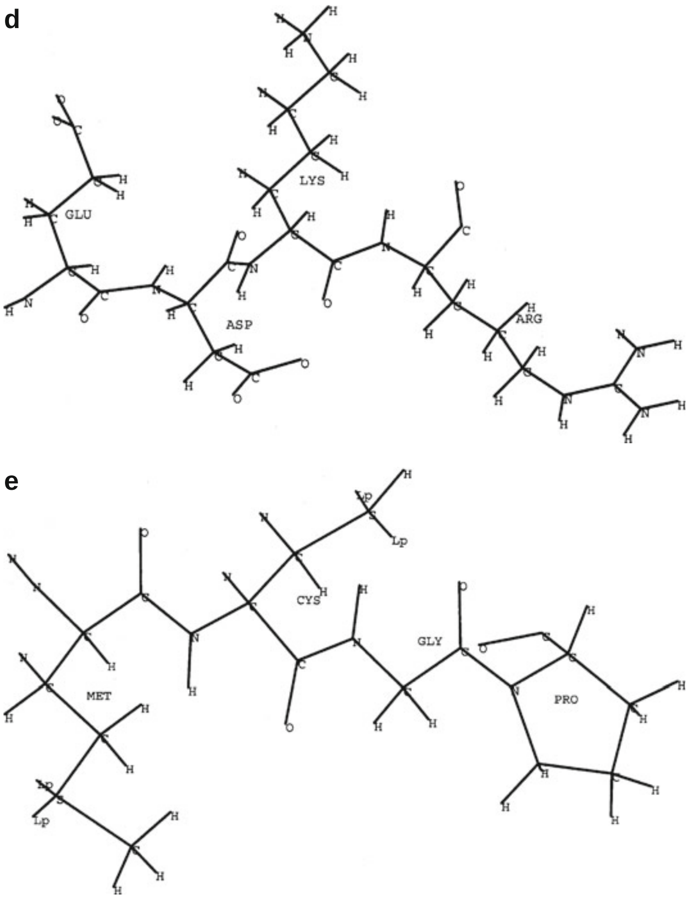


Fig. 18.10 (continued) The structures of the major biological amino acids are shown here as tetrapeptides. Each side chain is shown trans to the peptidyl backbone and the chemical connectivity can be ascertained in the framework model. A space-filling model then shows the relationship of the side chains to the backbone. (**a**) aliphatic: alanine, valine, leucine, isoleucine; (**b**) polar: serine, threonine, asparagine, glutamine; (**c**) aromatic: tryptophan, tyrosine, phenylalanine, histidine; (**d**) ionizable: glutamic acid, aspartic acid, lysine, arginine; (**e**) special: methionine, cysteine, glycine, proline

change the moiety from charged to uncharged. Because of this property, histidine is often found at active sites of proteins that show a sensitivity to pH. Histidine can be used as a proton relay because of the pK_a of the side chain.

Besides the energy associated with favorable entropy (hydrophobicity) and ion–*X* interactions, other interactional energies participate in side chain chemistry. These energies can be better appreciated by using a classification (Table 18.8) that reflects the chemical nature of the side chains. The aliphatic side chains are hydrophobic but can also demonstrate van der Waals interactions. The aromatic side chains can

Table 18.6 ΔG of transfer from ethanol to water and water to 8 M urea at 298 K

Amino acid	ΔG_t (ethanol to water) (J/mol/deg)	ΔG_t (water to urea) (J/mol/deg)
Glycine	−19.4	+0.42
Glycine (normalized)	0.0	0.0
Serine	−1.3	
Threonine	+1.7	
Alanine	+2.1	−0.29
Histidine	+2.1	
Methionine	+5.4	
Valine	+6.3	
Leucine	+7.5	−1.59
Tyrosine	+9.6	−3.05
Phenylalanine	+10.5	−2.93
Proline	+10.9	
Tryptophan	+14.5	

Data from Tanford C. (1962) *J. Am. Chem. Soc.* 84:4240 and Whitney P.L. and Tanford C. (1962) *J. Biol. Chem.* 237:1735

In order to determine the contribution from the side chain alone, glycine, which has only a hydrogen as a side chain, is used as the standard for comparison. By subtracting the $\Delta G_{\text{transfer}}$ of glycine from all other amino acids, the resulting value is the contribution to the energy of the side chain alone

Table 18.7 Acid–base properties of the amino acids

Amino acid	pK_a (α-COOH)	pK_a (α-NH3)	pK_a (side chain)	pI
Alanine	2.34	9.69		6.00
Arginine	2.17	9.04	12.48	10.76
Asparagine	2.02	8.80		5.41
Aspartic acid	1.88	9.60	3.65	2.77
Cysteine	1.96	10.28	8.18	5.07
Glutamine	2.17	9.13		5.65
Glutamic acid	2.19	9.67	4.25	3.22
Glycine	2.34	9.60		5.97
Histidine	1.82	9.17	6.00	7.59
Isoleucine	2.36	9.60		6.02
Leucine	2.36	9.60		5.98
Lysine	2.18	8.95	10.53	9.74
Methionine	2.28	9.21		5.74
Phenylalanine	1.83	9.13		5.48
Proline	1.99	10.60		6.30
Serine	2.21	9.15		5.68
Threonine	2.09	9.10		5.60
Tryptophan	2.83	9.39		5.89
Tyrosine	2.20	9.11	10.07	5.66
Valine	2.32	9.62		5.96

Table 18.8 The classification of amino acid side chains can reflect several aspects of their observable state space

Classification	Amino acid
Nonpolar (hydrophobic)	Alanine, leucine, isoleucine, valine, proline, phenylalanine, tryptophan, methionine
Uncharged polar	Serine, threonine, tyrosine, asparagine, glutamine, cysteine
Positively charged (pH = 6)	Lysine, arginine, histidine
Negatively charged (pH = 6)	Aspartic acid, glutamic acid
Hydrogen bond formers	
Acceptors and donors	Serine, threonine, asparagine, glutamine, cysteine
Donors only	Tryptophan, arginine
pH dependent	Lysine, aspartic acid, glutamic acid, tyrosine, histidine
Aliphatic	Alanine, valine, leucine, isoleucine
Aromatic	Phenylalanine, tyrosine, tryptophan
Aliphatic hydoxyl	Serine, threonine
Sulfur containing	Cysteine, methionine
Amide containing	Glutamine, asparagine
Special geometries	Glycine, proline

undergo charge transfer bonding. Hydrogen bonding occurs with the side chains of the alcoholic amino acids threonine and serine, the sulfhydryl of cysteine, the phenolic group of tyrosine, and the carboxamides asparagine and glutamine.

The sulfhydryl group of the amino acid cysteine deserves special attention because it has a series of interactions of varying importance. The pK_a of the cysteine side chain is 8.33 making it weakly acidic, and so while dissociation is not favored at physiological pH like histidine, local pH changes could have s significant effect on the ratio of dissociated to undissociated sulfhydryl. The side chain is capable of hydrogen bonding though weakly. Most importantly, at neutral pH, two reduced sulfhydryl groups in a pair of cysteine side chains can undergo oxidation and form a covalent disulfide linkage. These disulfide linkages play a crucial role in the tertiary structure of proteins and in a protein's biological activity.

18.5 Nonpolar Polypeptides in Solution

A model system of a macromolecule based on the properties of the constituent smaller residues can be built using thermodynamic information such as that listed in Table 18.6. The side chains of the nonpolar amino acids can be considered to behave as if they were small nonpolar solutes. Just as small molecules form stable hydrates in dilute aqueous solutions, the nonpolar polymerized amino acids can also form stable hydrates, which are much larger. The water molecules forming these hydrates surround the entire polymer, and the resultant structures are named *hydrotactoids*. Water in these hydrotactoids has a very different structure from water surrounding the small nonpolar molecules. The small nonpolar entities exist within the ordered H_2O structures as separate entities and act as isolated molecules. In contrast, the residues of nonpolar polypeptides have limited freedom of motion because the covalent peptide linkage imparts a structure to the polypeptide chain and therefore forces a non-random relative orientation on the nonpolar side chains. This forces the surrounding water molecules into even more rigidly held structures.

As already described, solvation of the small nonpolar side chains involves the rearrangement of the surrounding water in order to form pentagonal structures. The enthalpy change for solvation of the isolated nonpolar molecules is negative (exothermic). Incorporating these clathrate structures into the hydrotactoid results in a negative enthalpy of solvation for the side chain, but to a smaller degree than for the single molecules. The entropy change associated with the polymer solvation is highly negative, since a high degree of order is imposed upon the water molecules immediately adjacent to the polymer. The degrees of freedom for these water molecules are severely limited, since the number of possible microstates available becomes greatly reduced compared to that in the bulk water. The overall solvation of the polymer would surely be thermodynamically unfavorable ($\Delta G > 0$) if only the side chains were considered in the model; but most proteins, even highly hydrophobic ones such as tropoelastin, are soluble in aqueous solution. Possibly consideration of the constraints due to the peptide bond structure also needs to be incorporated into the model. Therefore, describing the relationship between the macromolecule and its aqueous environment depends on extending the model of non-ionic hydration for small independent molecules to the more restricted configuration of a macromolecule and its hydrotactoid. The ultimate configuration depends partly on interactions between the nonpolar residues, the sterically restricted rotational freedom of the polypeptide chain, and the structure of the water of solvation.

We have referred to the steric constraints on the polypeptide chain earlier. Normally, the planar peptide bond is a *trans* bond in polyamino acids. Careful molecular studies have shown that even the barriers to rotamer formation are relatively low; almost all polypeptides studied have shown that the side chain conformations are very near the rotational minima whether in the folded or unfolded state, therefore the *cis* form does not occur except in imino acid-containing polypeptides.

The periodic secondary structures are stabilized by hydrogen bonds, intrachain in the case of the helix and interchain in the case of the β-sheet (Fig. 18.11). A number of helical arrays, not just the more famous α-helix, are possible when external stabilization is present. The net effect of any of these ordered or periodic structures is to impose a higher degree of proximity on the nonpolar R groups. In fact, these nonpolar polymers generally make very compact helices in dilute aqueous solutions when there is no steric interference between the side chain R groups. A stable antiparallel pleated sheet structure is formed when the side chains sterically interfere with each other (Fig. 18.12).

These structures are the stable ones calculated on the basis of the polypeptide "backbone" chain. The more extended the polymer, the greater is the number of H_2O molecules whose state is perturbed. The thermodynamic result of this perturbation

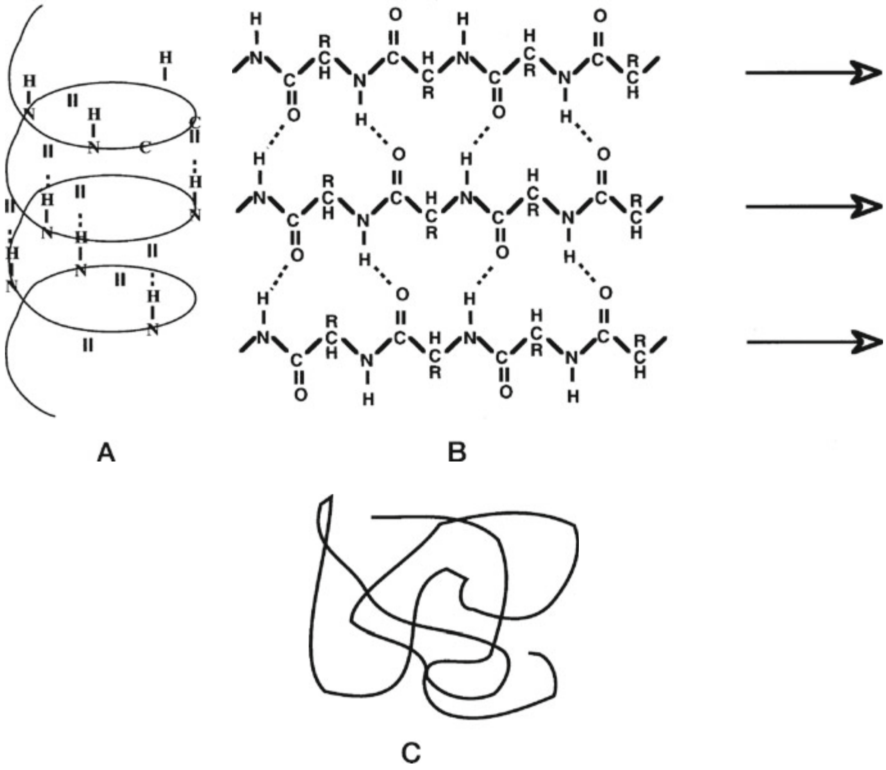


Fig. 18.11 The three available secondary structures for a polypeptide chain and their hydrogen bonding: (**a**) helix, (**b**) pleated sheet, (**c**) random chain (no periodic pattern)

depends upon the relative extent of interaction of the polymer with itself versus that with the aqueous solvent. If there is no steric hindrance, the polypeptide will fold so as to maximize the intrapolymer interactions and to expose the least surface to its aqueous surroundings. For example, a homopolymer of the nonpolar amino acid poly-L-alanine folds to yield an α-helix. In such a conformation, there are a maximum number of internal hydrogen bonds, because every peptide bond NH and C=O is hydrogen bonded. Because of the pitch of the helix, there is a rise of 3.6 amino acid residues for each full turn of helix. Each hydrogen bond is therefore made to the fourth peptide bond along the chain, so that all the hydrogen bonds align themselves parallel to the helix's axis, and the peptide backbone is fully hydrogen bonded. The exterior of this helix, which consists of the exposed side chains, is nonpolar. The helix is surrounded by H_2O molecules in their clathrate pentagonal structures, forming the hydrotactoid, but in a much more favored arrangement than that predicted for a long chain of unassociated nonpolar residues.

This behavior is observed in very dilute solutions of nonpolar homopolymers. However, in more concentrated solutions, the hydrotactoids will be close enough in

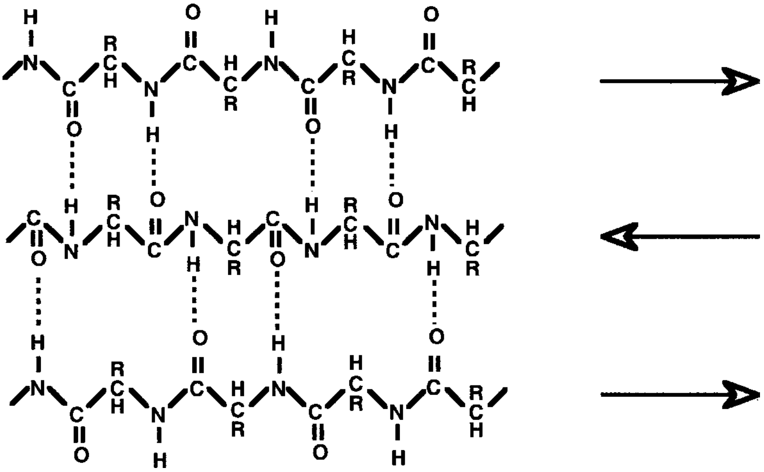


Fig. 18.12 A stable antiparallel pleated sheet prevents steric collision when R groups are large. The R groups will be located above and below the pleated sheet with hydrogen bonding occurring within the sheet

proximity so that removal of the H_2O molecules to the bulk water becomes possible. Under this condition, the overall entropy is increased. The polymer molecules associating so closely with each other will then no longer be adequately solvated and will aggregate and often precipitate (Fig. 18.13).

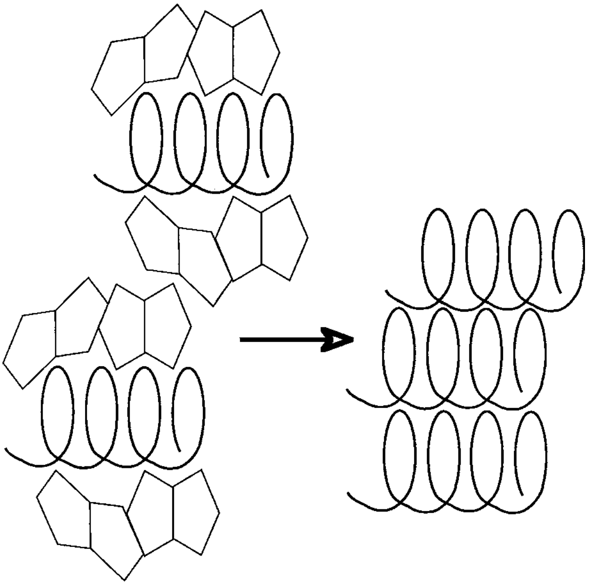


Fig. 18.13 At a high enough concentration of nonpolar polypeptide, the chains approach closely enough that the hydrotactoid water is removed to the entropically favored bulk water environment. Aggregation and precipitation will then occur

18.6 Polar Polypeptides in Solution

In contrast to the simple model of a nonpolar homopolymer (e.g., polyalanine) which is stabilized by intrachain hydrogen bonding, a polar homopolymer (e.g., polyserine) will be stabilized by hydrogen bonding with the aqueous milieu. If the R groups cannot ionize but are able to hydrogen bond with H_2O (as is the case for, e.g., polyserine), the polymer can take a helical conformation, and the external R groups will hydrogen bond to the aqueous solvent. An apparently normal hexagonally arrayed H_2O sheath can surround the molecule (Fig. 18.14).

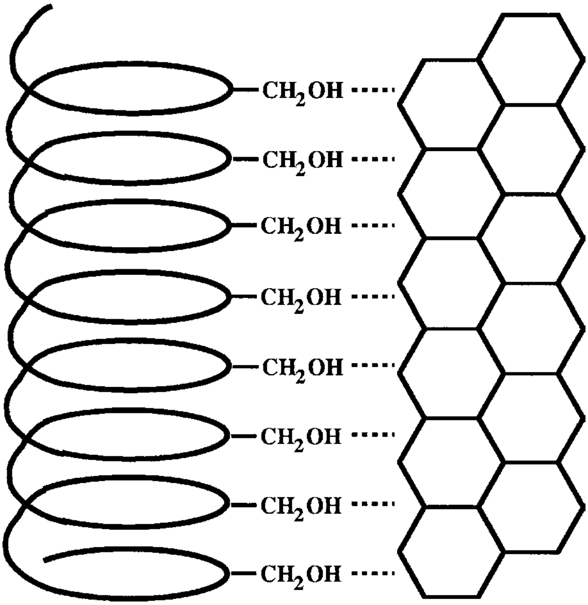


Fig. 18.14 A non-ionized polar polypeptide may form hydrogen bonds with the water structure surrounding it, with normal hexagonally arrayed water as the result

Finally, a homopolymer of an ionizable amino acid, e.g., poly-L-glutamic acid, can be considered (Fig. 18.15). The net charge on the polymer (the sum of the contributions from each side chain) will depend upon the degree of ionization of the side chains, which in turn depends upon the pH. If the polymer is unionized and consequently uncharged, the conformation can be helical and can under some circumstances (e.g., poly-L-glutamic acid at pH 3.0) be stabilized further by hydrogen bonds between the side chains analogously to polyserine. On the other hand, if the side chains of the polymer are ionized (e.g., poly-L-glutamic acid at pH 7.0), the identically charged side chains will repel each other and a helical structure cannot exist. Even a pleated sheet may not be stable. Instead, the homopolymer will assume the conformation in which the like charges are at the greatest possible distance from each other, forming a random chain configuration. The water structure around this

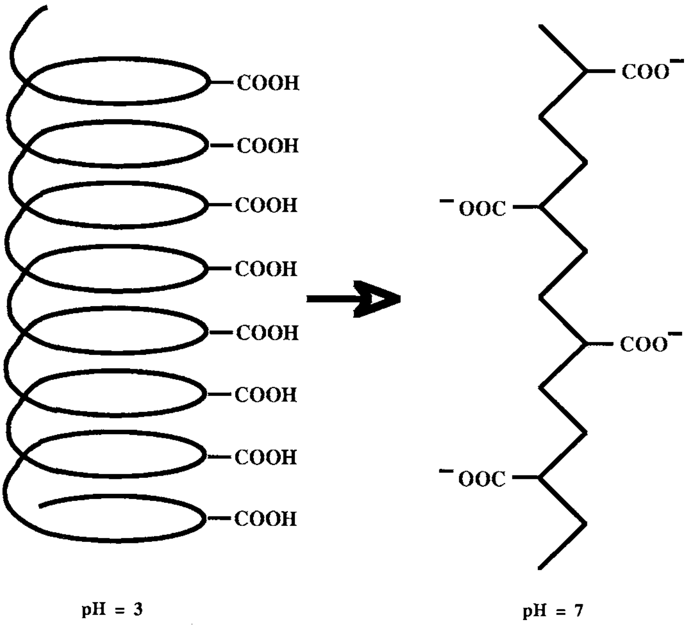


Fig. 18.15 Secondary structures that result as a consequence of the state of ionization of a polyelectrolyte such as polyglutamic acid

charged polymer will be altered. All hydrogen bonds between the molecules immediately adjacent to the polymer will be broken, and the hydration sheath will attempt to achieve the structure already discussed for small ions.

Biological polymers are rarely as uniform as the polymers discussed so far. Instead of a single repeating amino acid side chain, natural proteins have some hydrophobic, some polar unionized hydrophilic, and some charged residues. The protein will fold in such a manner that a minimum of energy will be required for its stabilization under the specific conditions of solvation (i.e., minimal ΔG). The actual sequence of these amino acid residues in any polypeptide chain as it is synthesized (i.e., before posttranslational modification) dictates the conformation of the polypeptide but the folding process is not at all obvious. This is generally referred to as “the protein folding problem.” The folded conformation is thought to maximize the extent of intrachain interactions that the given sequence of amino acids can accommodate. The forces operative in such folding are summarized in Fig. 18.16. The forces of interest include coulombic interactions between oppositely charged ions, ion–dipole interactions, hydrogen bonds, and hydrophobic interactions (such as van der Waals and London dispersion forces) when the entities are in virtual contact.

The overall folded structure is also influenced by the hydrophobic effect. It is important to note that hydrophobic effects are highly entropy-driven. Because the entropy of hydrogen-bonded bulk water is so high, H_2O molecules tend to

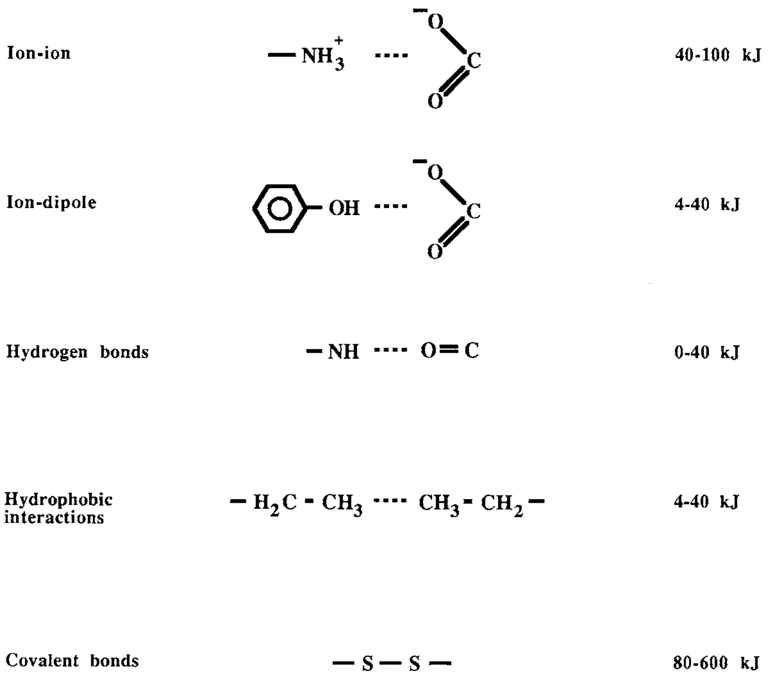


Fig. 18.16 Types of interactions involved in the folding of a macromolecule

be excluded from regions where they would be trapped. Molecules of water in a "hydrophobic pocket" have limited microstates available and do not hydrogen bond. Therefore, removal of these restricted H_2O molecules from a hydrophobic pocket leads to a significant increase in entropy.

In addition to the entropic gain in energy associated with the removal of water from the protein interior an enthalpic benefit also derives from this event. Many of the non-covalent interactions that lead to folding behavior are electrostatically derived and water is an important contributor to the local dielectric constant. The dielectric constant of the environment must be explicitly considered in understanding these interactions. The lower the dielectric constant, the stronger the interaction will be. The driving forces in folding a macromolecule in an aqueous environment ($\epsilon = 80$) will generally be toward making the region where folding occurs as anhydrous as possible, thereby decreasing the denominator in each of the force equations. X-ray investigations of pure protein crystals show very little trapped H_2O, generally less than three molecules per molecule of protein of molecular weight 14,000–100,000 Da. Thus, the driving force for the folding of macromolecules in aqueous solutions is strongly in a hydrophobic direction and seeks to minimize the number of water molecules located away from the hexagonally arrayed lattices of bulk water. Once the water molecules have been driven out of the region and the coulombic, ion–dipole, and hydrogen bonding associations are established, the stabilization energy for the folded structure will be high enough that the tendency of the macromolecule to retain this conformation will be very strong. Therefore, conformational transitions which allow the exposure of these regions of interactions to the aqueous milieu will not be favored. A description of the dependence of such transitions on the actual equilibrium conformation will be presented later.

Because conformational stabilization is achieved by the various interactions described earlier, there will be a strong dependence on the overall environment. In a homogeneous aqueous solution, the electrostatic and ionic interactions can be affected by the physical conditions at the time of structural analysis. For example, the electrostatic interactions between charged residues depend on the degree of ionization (i.e., on the pH), the ionic cloud of countercharge, and the local dielectric environment. In an aqueous system, ΔG is negative for an interaction between ionized residues of opposite charges, while ΔH and ΔS are both positive. Thus, the enthalpy change is unfavorable (energy must be expended to remove solvating H_2O molecules), but the entropy and therefore the free energy changes are highly favorable because the excluded H_2O molecules become part of the strongly hydrogen-bonded bulk H_2O. In the nonpolar interior of a macromolecule in its native conformation, these electrostatic or ionic interactions are strengthened by the low dielectric constant. In an aqueous environment, hydrogen bonds between residues or groups in the solute and the aqueous milieu will have a bond strength which approximately equals that of hydrogen bonds between H_2O molecules. Bonds between such groups do, however, become very much stronger in a nonpolar milieu. Finally, the van der Waals attractions that form the basis of the force of hydrophobic interactions are substantial. Once water has been excluded, it is highly unfavorable for the water to surmount the energy barrier to return to the nonpolar interior.

Even entities which possess quaternary structure, that is, which consist of several non-covalently bound subunits, remain strongly associated unless there is addition of some perturbing agent, such as a salt, a competing denaturant, a detergent, or a hydrogen-bonding entity; or unless the dilution is extreme. Only in very dilute solutions does a multimeric protein like hemoglobin dissociate into its monomeric subunits, while at physiologic concentrations these subunits are held together in the native multimer by strong hydrophobic interactions between the α and β subunits. In addition, electrostatic interactions (salt bridges) exist in this hydrophobic environment between the subunits, being stronger in the deoxygenated than in the oxygenated hemoglobin. Extraneous components or conditions (for example, the concentrations of ions or changes in temperature) can disturb the macromolecule's aqueous sheath and perturb the stable conformation of the biopolymer. These perturbing agents elicit changes in the macromolecular conformation of the solute and in its thermodynamic state. These transitions of state will be the topic of the next section.

The aqueous sheath is so highly ordered and conserved that the crystal structure of a protein formed in an environment where free water is not readily available (e.g., a high molar salt solution) is almost structurally indistinguishable from the structure of the protein in dilute aqueous solutions. This indicates that the formation of the aqueous sheath around a protein is probably highly favored and not simply a matter of the experimental technique involved in studying the protein structure. Since the amount or abundance of free bulk water in a cell is not clearly known, but is probably quite limited, this finding suggests that the structure of proteins produced in a cell is probably the same as the structure found by analysis from experiments done in dilute solutions.

The thermodynamic evaluation of solutions of macromolecules, especially of polyelectrolytes, is exceedingly complex even in simple dilute aqueous solutions. While there have been a number of theories attempting to explain the state of these solutions and to predict the nature of the aqueous medium which is their largest component, most of the theories have had limited success. Analyses are therefore performed experimentally, and the changes in thermodynamic state parameters, especially ΔS, are more readily determined than the absolute state constant of entropy.

18.7 Transitions of State

As stated earlier, biopolymers are always in a hydrated state, even in the crystalline form. The crystal structure of a protein can be shown to be identical to the conformation in solution under comparable conditions. Although the polymer may no longer contain the entire primary amino acid sequence present at synthesis, the conformation in which biopolymers are isolated has been defined as the *native* conformation. Although it is clearly a function of the conditions of isolation, including pH, ionic strength, and specific ions present, these parameters are frequently unspecified when the native conformation is described. The forces operative in stabilizing this native conformation include ionic, ion–dipole, and van der Waals forces, expressed as ionic and hydrogen bonds and hydrophobic interactions. All take place in as water-free an environment as the macromolecule can achieve, so that the interior dielectric constant can be kept low. Crystallographic evidence points to a very small number of internal “trapped” H_2O molecules in most proteins. Thus, while the hydration sheath is known to exist, its extent and properties are far from clear, although it is apparent that solvation plays a critical role in the stabilization of polyelectrolyte conformations. The greater the extent of intrapolymer interactions, the more native the conformation will be, while predominant polyelectrolyte–solvent interactions would favor an extended random conformation. Changes in the native state of a biopolymer yield a sizable amount of thermodynamic information on the energy and entropy changes involved in the transition, as well as an overview of the type, strength, and interdependence or *cooperativity* of the stabilizing interactions in each state.

In describing the conformational stages that a macromolecule goes through as it changes from a native to a denatured state, the simplest model for the change in state which can be achieved by a change in temperature, pH, ionic composition, or solvent can be represented as a series of steps from the native state, *n*, through states $y_1, y_2, y_{3,} \ldots$, to the denatured state, *d*:

$$n \rightarrow y_1 \rightarrow y_2 \rightarrow y_3 \rightarrow\rightarrow\rightarrow d \tag{18.74}$$

At any point during the transition the total system can be represented as the sum of the individual component states, that is, as the sum of the fractional contributions from each state. This is written in terms of the mole fractions as the sum of the mole fraction X_x times the state $y_{x,}$ where y_n is the native state, y_d is the denatured state, and y_i are the intermediate states. This can be written as

$$1 = X_n y_n + \sum_i X_\mathrm{i} y_\mathrm{i} + X_d y_d \tag{18.75}$$

For each step in the denaturation, there is a different equilibrium constant $K = X_i/X_{i\text{–}1}$, which is the ratio of the mole fraction of state *i* divided by the mole fraction of the previous state, *i* – l. The denaturation curve for such a system might look like Fig. 18.17. The abscissa represents the magnitude of any given perturbation which brings about the change in state, and the ordinate shows the ratio of the native to the denatured state for a macromolecule.

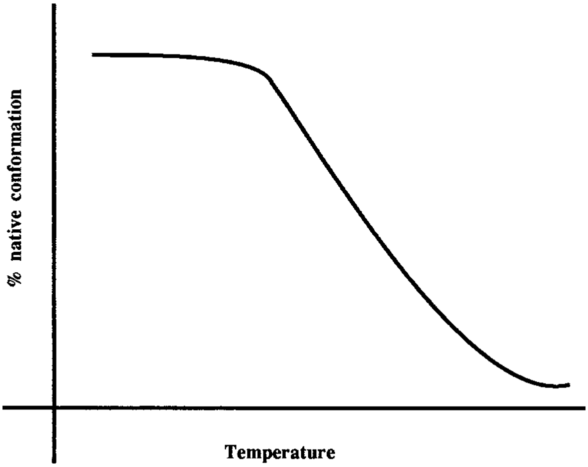


Fig. 18.17 Sample denaturation curve

Most actual denaturation curves follow this pattern, which is asymmetric about its midpoint halfway between the native and the denatured states. This type of curve indicates that there are a number of intermediate quasi-stable, although not necessarily isolable, states between the native and the denatured end points. Mathematical treatment of such asymmetric curves, and the extraction from them of the free energy and the entropy change associated with each change in state, is extremely complex and often impossible unless the intermediate states themselves can be characterized. Therefore, the discussion will be limited to the special case in which only two stable states, the native, *n*, and the denatured, *d*, can exist. In this special case, at every intermediate point between 100% native and 100% denatured, a simple mixture of X_n and X_d in various proportions exists and accounts for all the material:

$$X_n + X_d = 1 \tag{18.76}$$

X_n is the mole fraction that is native and X_d is the mole fraction in the denatured state. Then, since

$$y_n \Leftrightarrow y_d \tag{18.77}$$

and, at any given quantity of the perturbing agent, the total, y_t, is

$$y_t = X_n n + (1 - X_n)d \tag{18.78}$$

where X_n is the fraction remaining in the native state, then at any chosen value of the parameter causing the change in state (e.g., *T*), the ratio of native to denatured states, f_T, can be represented as

$$f_T = \frac{X_n}{1 - X_n} \tag{18.79}$$

There is just one equilibrium constant:

$$K = \frac{(X_\text{n})_\text{eq}}{(1 - X_\text{n})_\text{eq}} \tag{18.80}$$

Such a denaturation can be represented by Fig. 18.18.

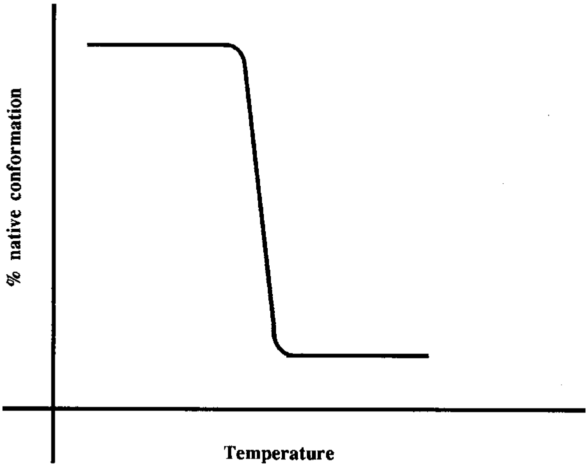


Fig. 18.18 Denaturation curve for the special two-state case

In this two-state-case system, the curve must be fully symmetric about the midpoint at $f = 1.0$; this is the easiest way to recognize a system in which only the native and the denatured states can exist. The curves can indicate not only the presence of two states, but also the *cooperativity* inherent in the change from one to the other. Cooperativity is a measure of the increased likelihood that once a given residue in the polymer has undergone a conformational change associated with denaturation, the next residue along the chain will also undergo such a change. Graphically, cooperativity is represented by the slope of the denaturation curve at its midpoint. A parameter, s, was defined by Zimm as the probability that once a residue is in a given conformation (e.g., denatured) the next one will also be in the same conformation. Therefore, s is a measure of the interdependence of the conformations of adjacent residues. If these conformations are independent, the probability s will be 1/2; if totally interdependent, s will be 1. A second parameter, σ, is a measure of the ease with which a residue exhibits a conformation different from that of its predecessor, that is, of the ease of initiating a change in conformation. Therefore, σ will be 1 when there is no restriction on such an initiation, and it will be 0 when there is no chance of an initiation, that is, when the entire molecule *must* be in the same conformation.

The equilibrium constant for the nucleation or start of the conformational change can be defined as σs:

$$\sigma s = \frac{...rrrrhrrrr}{...rrrrrrrrr} \tag{18.81}$$

where r is a polypeptide residue in the random configuration and h is one in the helical configuration. This could also be written as

$$\sigma s = \frac{\text{Single change in conformation}}{\text{No change}} \tag{18.82}$$

The constant s then represents the likelihood that once a region is started in a given conformation, it will continue in that conformation (a “zipper” effect), exhibiting cooperativity or interdependence of components:

$$s = \frac{...rrrrhhhhhrrrr...}{...rrrrhhhhrrrrr...} \tag{18.83}$$

That is,

$$s = \frac{\text{Propagated}}{\text{Non propagated}} \tag{18.84}$$

If s is high, there is extensive interdependence. The likelihood that the next residue will also undergo denaturation will be greater the higher is the value of s. The smaller the σ, the fewer is the possible number of locations where initiation of denaturation will start. This type of cooperativity accounts for the steepness of most

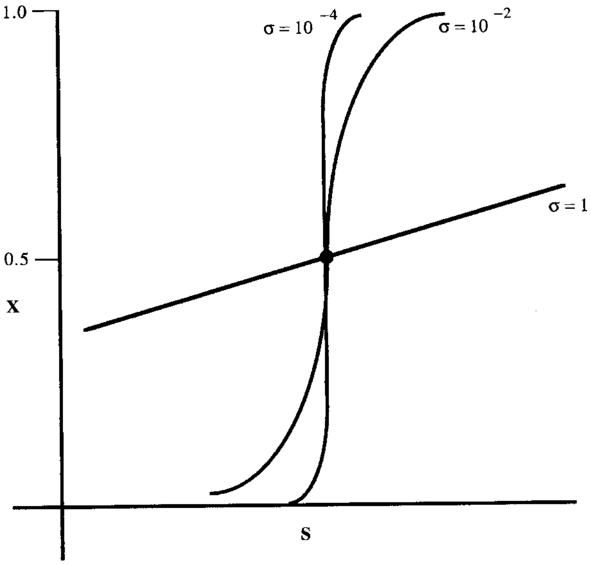


Fig. 18.19 Family of curves showing the relationship between *s* and *σ*

biopolymer or protein denaturations when the overall conformation is observed (Fig. 18.19).

When a protein such as ribonuclease is denatured by heat, the appearance of the denaturation curve and the cooperativity of the process depend upon the conditions employed and the parameter of denaturation which is chosen to be measured. If the parameter reflecting the extent of native structure is characteristic of the entire molecule, which intrinsically means that it is less sensitive to a change in conformation involving only a few (or a small percentage) of the residues, the apparent cooperativity will be great. Ribonuclease denaturation will exhibit strong cooperativity when it is measured by a circular dichroism change in the peptide bond absorbance region around 200 nm (Fig. 18.20). Under the same conditions, if the parameter observed is characteristic of only a few residues, for example, for ribonuclease the circular dichroism in the region of aromatic side chain absorbance, then the denaturation curve is no longer either symmetric or highly cooperative (Fig. 18.20).

In an aqueous system, a biopolymer is surrounded by a hydration shell, but in the native state it contains very few internal H_2O molecules. A change in conformation by a single residue in such a biopolymer may permit the insertion of H_2O into what had been a water-free environment. Since the internal interactions stabilizing the native conformation are all weakened in the presence of H_2O, the next residue will also become denatured, then the next, and so on. Therefore, denaturation of a biopolymer in an aqueous medium is likely to be accompanied by a high value of *s* and is highly cooperative. In the example of a biopolymer such as DNA, the steepness of the denaturation curve, *s*, will be a measure of the homogeneity of the

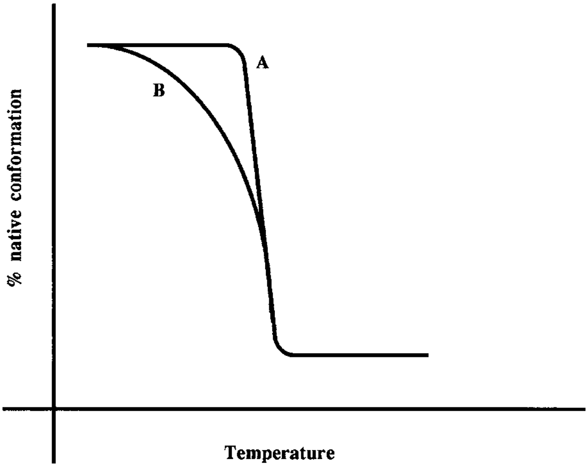


Fig. 18.20 Thermal transition curve obtained using circular dichroism measurements that show the differential effect of measuring a parameter representative of all the peptide bonds (curve A) versus one that examines only a small proportion (1/6) of a smaller population (6/123) that is sensitive to a conformational change (curve B)

DNA and of the existence of well-hybridized double-helical regions, rather than a mixture of partially single-stranded or poorly hybridized species. Poorly hybridized species exhibit a higher value of σ, since the base pairing will be partially imperfect in poorly hybridized specimens. These facts form the basis for the pragmatic curves by which DNA is often characterized.

Thermodynamic values can be determined for any denaturation, those of greatest interest being ΔG, ΔH, and ΔS for thermally induced changes of state. If the process can be approximated by a two-state model with the fraction of molecules in the native state being represented by X_n and that in the denatured state by X_d, then

$$X_\mathrm{n} + X_\mathrm{d} = 1 \tag{18.85}$$

Then for ΔG for a state that is not the standard state

$$\Delta G = \Delta G^o + RT \ln \frac{X_\mathrm{d}}{X_\mathrm{n}} = -RT \ln \left(\frac{X_\mathrm{d}}{X_\mathrm{n}}\right)^\mathrm{o} + RT \ln \frac{X_\mathrm{d}}{X_\mathrm{n}} \tag{18.86}$$

A term K_app can be defined:

$$K_\mathrm{app} = \left(\frac{X_\mathrm{d}}{X_\mathrm{n}}\right)^\mathrm{o} \left(\frac{X_\mathrm{d}}{X_\mathrm{n}}\right) \tag{18.87}$$

which gives the result

$$\Delta G = -RT \ln \left(\frac{X_\mathrm{d}}{X_\mathrm{n}}\right)^\mathrm{o} \left(\frac{X_\mathrm{d}}{X_\mathrm{n}}\right) = -RT \ln K_\mathrm{app} \tag{18.88}$$

But $\Delta G = \Delta H - T\Delta S$, and therefore

$$\Delta \frac{H}{T} = -R \ln K_{\mathrm{app}} + \Delta S \tag{18.89}$$

Equation (18.89) is the *van't Hoff equation*. If it can be assumed that ΔH and ΔS are independent of T, a plot of ln K_{app} versus $1/T$ (in Kelvin) should yield a straight line whose slope is $-\Delta H/R$. If the assumption that only two states (native and denatured) exist is valid, the change in enthalpy at the midpoint of the transition, T_{m}, when $K_{\mathrm{app}} = 1$ can be calculated from the slope of the denaturation curve at that inflection point. The enthalpy change obtained by application of the van't Hoff equation is called $\Delta H_{\text{van't Hoff}}$. The change in enthalpy derived in this way is valid only for the particular conditions of pH, salt concentrations, etc., under which the transition was measured but is nevertheless a useful parameter in the comparison of the denaturation of polyelectrolytes. In general, the most direct way to obtain ΔH is by calorimetry, but such measurements on macromolecules have proven difficult to perform.

If the assumption of a temperature-independent ΔH is not valid, the relationship between ΔH, ΔS, and K_{app} can nevertheless be used to determine ΔS for the denaturation at various temperatures. It will generally be positive, since the polymer chain has more available configurations than a more tightly held periodic structure. Although the random form may require a less favorable solvation sheath structure, thereby reducing the overall system entropy, unfavorably solvated random chains will lead to aggregation of the denatured species in an attempt to reduce the aqueous sheath. The overall ΔS for denaturation therefore remains positive.

We have concentrated on thermally induced changes of state, because the thermodynamic parameters ΔH, ΔG, and ΔS are more readily understood in this context. It should however be noted that denaturation or change of phase can be brought about by any of a number of perturbations other than heat, and the same kind of thermodynamic analysis can be applied. Using ribonuclease as an example of a typical system, the curve for the change in state from the native to the denatured state as a function of added urea concentration can be plotted. Denaturation by addition of urea yields a curve which can be analyzed in exactly the same way as thermal denaturations in terms of deducing ΔH, ΔG, and ΔS of the denaturation as a function of urea concentration rather than as a function of temperature. The higher the concentration of urea at the midpoint of the curve, the stronger the interactions within the native conformation of the solute will be. The steeper the denaturation curve, the more cooperative the denaturation will be once denaturant has begun to compete successfully for hydrogen bonds outside or within the protein. If the curve is symmetric about its midpoint at 50% native conformation, the system has only two possible conformations, a native and a denatured one, and only the approximations valid for a two-state system may be applied.

18.8 The Protein Folding Problem

Though other biopolymers exist in nature, none have the wide-ranging variability of polypeptides and proteins. The issues of predicting the folding path of proteins have been fundamental in the last half of the century and continue to be an extremely important area of research, especially today when we have the ability to produce many primary structures at will via molecular biological and biotechnology approaches. Since C. Anfinsen demonstrated (Fig. 18.21) that ribonuclease could be folded and unfolded into its native (*active*) three-dimensional structure as long as the primary structure was intact, it has generally been accepted that the primary sequence contains sufficient information to direct the complete native folding of a protein. In the decades prior to this demonstration the observation had been made that a protein could be denatured by treatment with heat, acidification, or chaotrope with occasional and usually unpredictable renaturation. With the supposition that the pattern of nucleic acids leads to the pattern of proteins (DNA → RNA → protein), which represented the one gene–one protein theory, Anfinsen's experiments linked this molecular biological theory to the denaturation observations. Ample experience has shown that simply the production of a correct primary structure does not assure active or useful configuration. A renewed interest, with substantial economic implications, in learning how to make active configurations of proteins from the primary genetic information is now upon us.

In spite of the difficulties implied by the preceding sentences, it is probably generally correct to view the linkage between primary structure and folded–functional protein as one of chemistry; with the proper adjustment of solvent and conditions, the proper folding pathway should lead to active protein. It is known that other factors, those of biological chemistry, also play a crucial role in certain folding paradigms. These factors include the disulfide exchange enzymes, proline *cis*–*trans* isomerases, and the chaperonins. In certain cases these other "*bio*chemical" agents may have to be added to the state space in which a protein folds in order that error and surprise are not experienced.

In its simplest form the protein folding path can be written as

$$\text{Unfolded} \quad \mathbf{U} \longleftrightarrow \mathbf{N} \quad \text{Folded} \tag{18.90}$$

There will be a rate constant (first order) that describes the interconversion of the two states (U, N). Since a polymer contains many segment units it is useful to consider that a large number of related structures actually exist in equilibrium with Eq. (18.90).

$$\begin{array}{ccc} \mathbf{U_1} & \longleftrightarrow & \mathbf{N}_1 \\ \uparrow\downarrow & & \uparrow\downarrow \\ \mathbf{U_2} & \longleftrightarrow & \mathbf{N_2} \\ \uparrow\downarrow & & \uparrow\downarrow \\ \mathbf{U_3} & \longleftrightarrow & \mathbf{N_3} \\ \uparrow\downarrow & & \uparrow\downarrow \\ \uparrow\downarrow & & \uparrow\downarrow \\ \mathbf{U}_{\boldsymbol{n}} & \longleftrightarrow & \mathbf{N}_{\boldsymbol{n}} \end{array} \tag{18.91}$$

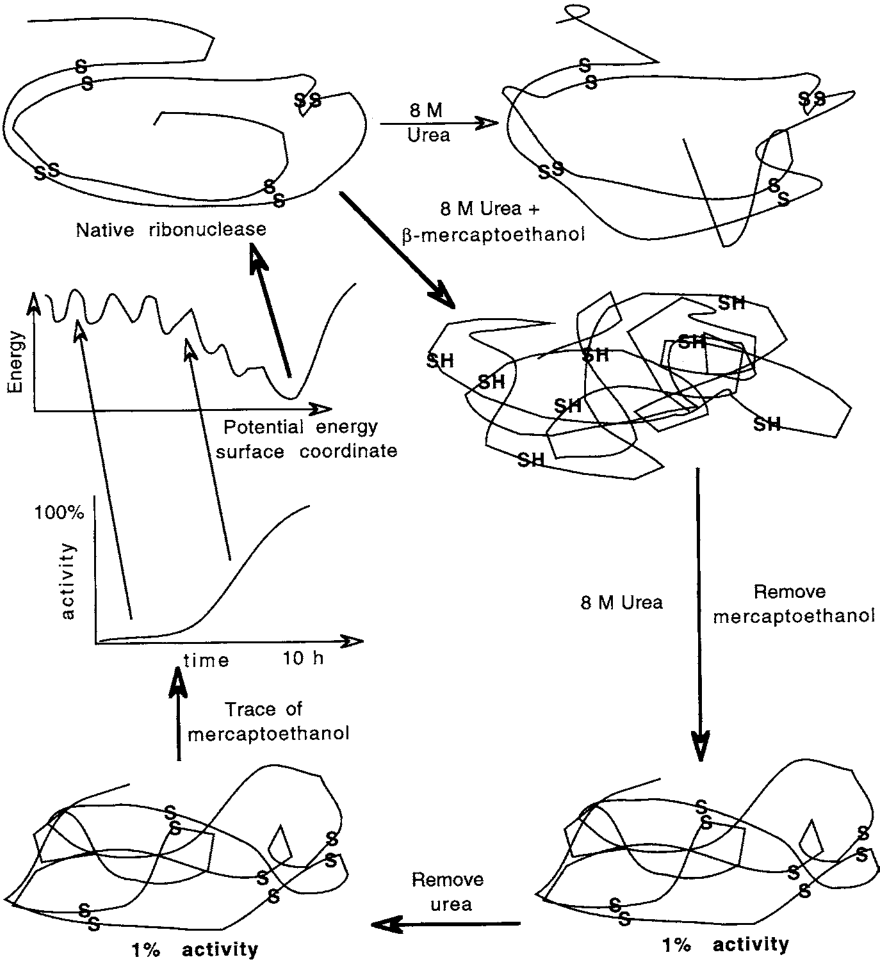


Fig. 18.21 Anfinsen's ribonuclease experiment. *Top*: The addition of a denaturant, urea, did not lead to full loss of tertiary structure in the absence of a reducing agent. *Counterclockwise*: The addition of reducing agents and urea leads to no activity and complete denaturation of the protein. If only the reducing agent is removed and the protein was allowed to air oxidize, random disulfide bonds formed. This overall mixture recovered only 1% of its activity. Since there are 105 ways of forming 8 sulfhydryls into 4 disulfide bonds and only one of these permutations is correct this supports a random formation of disulfides. Removal of the urea and the addition of a small amount of reductant allowed the shuffling of the disulfides until the correct native structure was spontaneously recovered. This process took approximately 10 h with air oxidation and was driven by the free energy as the enzyme moved down the potential energy surface to the stable minimum, which was the native structure

If a rapid equilibrium exists among the forms of U and also among the forms of N while the rate constant between the U and N states is relatively slow, the extensive equation shown in (18.91) will reduce to the case of Eq. (18.90). At equilibrium the process may appear two-state, but kinetic experiments can be expected to show a variety of states and rate constants. Alternatively multiple intermediate steps may be found between U and N leading to a much more complicated array of possible reactions:

$$\begin{array}{ccccccccc}
\mathbf{U_1} & \longleftrightarrow & \mathbf{I_1} & \longleftrightarrow & \mathbf{J_1} & \longleftrightarrow & \mathbf{K_1} & \longleftrightarrow & \mathbf{N_1} \\
\uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow \\
\mathbf{U_2} & \longleftrightarrow & \mathbf{I_2} & \longleftrightarrow & \mathbf{J_2} & \longleftrightarrow & \mathbf{K_2} & \longleftrightarrow & \mathbf{N_2} \\
\uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow \\
\uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow & & \uparrow\downarrow \\
\mathbf{U}_{\boldsymbol{n}} & \leftrightarrow & \mathbf{I}_{\boldsymbol{n}} & \leftrightarrow & \mathbf{J}_{\boldsymbol{n}} & \leftrightarrow & \mathbf{K}_{\boldsymbol{n}} & \leftrightarrow & \mathbf{N}_{\boldsymbol{n}}
\end{array} \tag{18.92}$$

The similarities between the two models of allosteric binding presented earlier and these two models of state changes should be noted. The first model is essentially the concerted model, while the second represents the sequential treatment model. The same statistical analysis that we performed earlier for multiple binding may be applied to this problem. A variety of models for protein folding have been proposed and include the sequential, nucleation/growth, framework, hydrophobic collapse, and jigsaw puzzle models. Each of these models can be described by manipulation of the rate constants in Eq. (18.92). These general models for protein folding cluster in one of two limiting case models. The sequential model is a linear, step-by-step model in which the rate constants of the top row are much greater than any other rate constants and so there is a linear movement along a structural state space from U to I to J to K to L to N with one isolable structure per step. The diffusion–collision and framework models are subtypes of this linear model. In both of these models a variety of intermediates can be shown to exist and the amount of any intermediate is a consequence of the ratio of a series of rate constants. The movement along the pathway to native structure is *continuous* though the parallel paths give a certain complexity to the process. In the diffusion–collision–adhesion model intermediates exist and diffuse together with a resulting collision. The effectiveness of a collision in terms of moving along the potential energy surface is high and the rate limiting steps for folding is considered to be the diffusion process that brings the intermediates together. The rate limiting constants will reflect diffusion and subsequent rate constants reflect the stabilization process through adhesive interactions. The framework model is similar in terms of its continuous nature but the intermediates are considered as formed portions of the final stable conformation. The time to bring the intermediates together is less important than the process of finding the correct docking geometry that leads down the path to a stable folded molecule.

The other type of model can be considered as having more discontinuous features than those of the sequential family. These models have some of the features of the catastrophe systems that we have discussed earlier and can be expected to show a sudden collapse into the native structure once a certain point is reached on the potential energy surface. The limiting model for this discontinuous family is the nucleation-growth model which can be modeled by making all of the rate constants going from U to I very slow and rate limiting and then having all other steps extremely fast. There will be no detectable intermediates once the folding process starts but due to the slow first step there may be the sudden appearance of many small transiently stable nuclei (these are most likely kinetic starting points or kinetic nuclei as opposed to the actual structural nuclei seen in crystal growth) followed by a rapid movement to the final structure. Not surprisingly there has been little experimental evidence to support such a model given the increasing number of systems that have been shown to have intermediates associated with their folding. However, catastrophe surfaces have continuous and discontinuous features. We could expect that a more realistic model would include a certain continuous movement to a point where a general structure might appear, driven into being by a dominant force and associated rate constants. The production and appearance come with the movement of the molecule over the catastrophe cusp. Subsequent movement and rate constants will reflect the local reorganization of the dominant intermediate structure into the final native structure. This (loosely) formal description is reflected in the hydrophobic collapse model. In this model the hydrophobic effect is considered to be the driving force taking a somewhat disorganized folding mass and moving inexorably toward a state of maximal dehydration. At some point a single structure appears that foretells the ultimate native structure. With a little more movement along the hydrophobic force lines, the secondary and tertiary structures appear and are stabilized. This model also has many of the qualities of the molten globule state, a model that is accumulating experiential evidence.

The process of folding a protein is best appreciated as occurring on a complex potential energy surface. Given the many local minima (3^{100}) for our original peptide it has been pointed out that a complete set of independent trials each occurring in less than a tenth of a picosecond could require over 10^{27} years to find a structure. Thus the folding process must retain folding steps that are correct and will constrain the number of possibilities reducing the overall process to one of real-time kinetic likelihood. Another aspect of the problem is that there are a large number of thermally accessible states near the native structure of proteins (Elber and Karplus estimated 2000 for myoglobin) and probably near any important structural intermediate. Thus the folding process will be very fluid and will continually move among these accessible states as it finds the region of the stable minimum. There is accumulating evidence that proteins can be shown to form locally stable secondary structures as we have detailed earlier. These locally energetically minimized structures then form a seething mass of relatively stable local domains that proceed in parallel toward the final overall structure. This state is aptly named the molten globule state. Some evidence exists to suggest that the molten globule state is formed by the hydrophobic collapse of the local secondary structures. This is an area of research still richer in models than in answers and the final solution will likely have aspects of each of the models discussed. We can probably expect a surprise or two as well.

18.9 Pathological Protein Folding

At this point we accept that the primary sequence contains the information for proper protein folding but recognize that while necessary this observable may not be sufficient to ensure a functional protein. While Anfinsen's ribonuclease refolded after denaturation, anyone who has ever cooked an egg knows that once denatured not all systems will functionally refold even if the primary sequence remains. This problem may be termed the *protein misfolding* question.

Work by M. Goldberg in the 1970s provides a perspective on this problem. In an effort to study the refolding of chymotrypsin, the protein was denatured and then allowed to refold in a test tube. Initial studies found that the protein did not functionally refold but instead aggregated into clumps. Importantly the clumps formed following an initial phase in which proper protein refolding was initiated. In the process of refolding, residues and secondary structures apparently came into contact and a new pathway, away from functional protein, was followed by the chymotrypsin. Goldberg found that by lowering the concentration of the refolding mixture, the yield of properly folded protein could be increased. The important interactions leading to misfolding were those between the protein itself. As we will see in the coming study of chemical and enzyme kinetics this problem of picking the correct path out of a series of available paths is a kinetic one.

A similar case to the high-concentration conditions that Goldberg employed exists in the endoplasmic reticulum where the nascent unfolded protein must undergo folding. While optimal in vitro folding occurs at concentrations of approximately 1 mg/ml, the protein concentration in the endoplasmic reticulum is of the order of 200 mg/ml. Furthermore the primary structure chain of a protein is fed into the ER C-terminal first followed by the N-terminal seconds to minutes later. How are the incorrect folding paths limited so that the cell can maintain a high yield of correctly folded protein? There are a variety of proteins in the lumen of the ER called *chaperone proteins* that briefly bind the chain of peptides and reduce the chance for entanglement that will lead to misfolded protein. There are a variety of chaperone proteins, the most common being a 78-kDa heat shock family class protein called BiP (Binding Protein). These heat shock proteins are induced when a cell undergoes severe stress and will start synthesizing a large number of new proteins in response to that stress. Thus, under normal conditions biological systems have evolved a system to ensure that the correct path (over a potential energy surface) is followed. This greatly increases the engineering efficiency of the cell. This technique of maximizing the kinetic yield of the correct product will be illustrated again in our study of enzyme catalysis. Improper paths can also be limited by restricting the choices a polymer chain can make in its random walk. Two proteins in the ER perform this function, protein disulfide isomerase which catalyses the formation of the correct disulfide bonds in a chain and peptidyl prolyl isomerase which influences the kinetics of protein folding by accelerating the *cis–trans* isomerization of a proline containing dipeptide (X-Pro).

Misfolding of a protein can occur because of interaction between secondary structure subunits; we have learned that the interactional energies holding those units in place are in general quite small. Thus the linkages in state space for misfolding can be simply the appropriate environmental change or a simple mutation in a protein that greatly influences the local secondary structure. A protein almost correctly folded but with a "loose" helix or dangling β-sheet may dynamically meet another similar protein and aggregate with that protein. Another possibility is that one protein with an abnormal dynamic structure can induce another protein with a normal structure to become dynamically unstable. Thus the abnormal protein can propagate or "infect" the normal protein. These pathways are available because the interactional energies leading to stable protein structure have low energy barriers and many relatively similar states are thermodynamically accessible on the potential energy surface.

This description of the misfolding process appears to be the common pathogenic pathway for the amyloid diseases and they provide a nice summary of the ideas that we have developed to this point.

The amyloidoses are a series of diseases that share the following observable: there is an extracellular accumulation of a substance that with histochemical staining has the microscopic appearance of starch, hence the name given by Virchow (*amyl* = starch, *oid* = like). However, amyloid is proteinaceous, not carbohydrate in nature. Amyloid is characterized by its staining with the dye, Congo red. When Congo red interacts with amyloid, the dye becomes characteristically apple-green birefringent when viewed under a polarizing microscope. Though amyloid is proteinaceous, it is characterized by its secondary and tertiary structures, not its primary structure. The overall configuration of all amyloids is that of a twisted fiber of defined proportion composed of peptides with a secondary structure of twisted β-sheets. It is likely due to the electronic interaction between the Congo red molecule and the β-fibrillar structure of the amyloid that gives rise to the diagnostic birefringence. While all the amyloids share secondary and some tertiary structural similarities, the primary structure is different in the different clinical diseases. The present view is that abnormal protein folding of the variety of proteins causing the amyloid diseases leads to the final common pathway of the accumulation of this dysfunctional residual and that either the deposition/accumulation of the amyloid material is lethal to the tissues or some toxic process associated with the amyloid or its deposition is the cause of the cellular pathology.

All amyloid fibrils, regardless of the biochemical or clinical type, share a typical composition at least when studied at the level of the electron microscope. Amyloid deposits are comprised of an array of linear, rigid, non-branching aggregated fibrils that are 7.5–10 nm in diameter but of indefinite length. The fibrils are formed from two or more filaments, each with a diameter of 0.25–0.35 nm, that form a helically twisted structure with a hollow central core. The helical structure gives the amyloid fibril a beaded appearance. The amyloid fibril substructure has been studied by x-ray diffraction techniques which show them to have a β structure. These observations have led to a model of amyloid fibrils in which antiparallel β-pleated sheets are the basic conformation. This model is supported by the extensive antiparallel β-pleated sheet structure found in the native structures of light chain immunoglobulins and transthyretin. The binding of the planar Congo red molecule is considered to be along the axial folds of the twisted β-sheet helical filaments.

Let us add a clinical dimension to the amyloid diseases.

18.9.1 Alzheimer's Disease

Dementia of the Alzheimer's type (DAT) is the most common degenerative disease of the brain in the Western world at the present time. Dementia is defined as the loss of already acquired mental capabilities. Over 4 million people in the United States are affected by DAT, and it has an incidence rate of approximately 123 cases per 100,000 in the general population. The incidence rate increases strongly with age. DAT is a slowly progressive dementing process in which the clinical disease has a relatively specific focal pattern of loss. These patterns of clinical degeneration are associated with focal accumulation of amyloid deposits in the portions of the brain responsible for the mental capabilities that are being ground away by the disease. The typical features of DAT are a triad of lost functions. The first is a loss of language function especially with common naming. This is a function of the dominant hemisphere of the brain. The second is a loss of ability to retain spatial orientation. These tasks are handled in the non-dominant side of the brain in locations, which are arranged as a mirror image of the language section of the dominant lobes. The cortical structures responsible for language and spatial orientation are those that subserve the analysis and integration of the special senses of the head. The third clinical feature of the triad is loss of memory, especially those which are newly formed. Formation of new memories is a function of midline structures that are anatomically and physiologically related to the "head" senses, namely the nuclei of the amygdala and the entorhinal and hippocampal cortices. In Alzheimer's disease all of these specialized areas are specifically involved as sites of substantial accumulation of amyloid substance and undergo specific atrophy and neuronal dysfunction and death. The relationship between the accumulation of the amyloid in deposits called amyloid plaques and the dysfunction of these areas is postulated to be causal. In fact the best that can be argued at our present level of knowledge is that the state space of amyloid accumulation and the state space of the neurological dysfunction in DAT have a high concordance.

The amyloid plaque found in brains afflicted with DAT is composed of a variety of proteins in addition to the principal protein, *amyloid β protein*. (This protein is alternatively called β/A4, βAP, or Aβ.) Other proteins associated with the plaque are α_1-antichymotrypsin, apolipoprotein E, substance P, and the tau protein. All of these proteins can be removed from the plaque using chemical denaturation techniques and leaving a core of Aβ. The residual Aβ is a mixture of peptides β1–42 and β1–43 which are polypeptides consisting of amino acids 1 (the N-terminal) through the 42nd or 43rd residue. Normally a shorter polypeptide β1–40 is found in the blood and cerebrospinal fluid, suggesting that soluble β-protein has a natural function. The 40, 42, and 43 residue polypeptides are created by cleavage from a transmembrane-bound portion of a larger protein, *amyloid precursor protein (AβPP)*. The formation of pathological insoluble amyloid seems to depend on the improper cleavage of AβPP into a 42- or 43-residue peptide instead of the natural Aβ1–40 peptide. This would suggest that the C-terminal end of the molecule plays an important role in the formation of insoluble amyloid fibrils from the soluble Aβ peptides. Modifying the natural C-terminus with the extra di- or tri-peptide might enhance the formation of the amyloid fibril. In terms of the chemical mechanics of fibril formation it would be expected that these C-terminal modifications would either enhance the stability of the forming fibril or increase the rate of fibril formation with respect to the rate of fibril denaturation. Either of these mechanisms can be expected to lead to accumulation of insoluble amyloid fibrils. It is interesting that the charged residues in Aβ are found in the β1–28 sequence while later C-terminal third contains primarily hydrophobic residues. Residues 42 and 43 are also hydrophobic which might be expected to enhance the tendency of the C-terminal portion to aggregate to avoid aqueous interaction. The formation of the fibrils is likely to depend on a variety of physical factors leading to nucleation of the peptides and accumulation of the peptidyl chains driven by favorable kinetics and thermodynamics. The death of the brain tissue results either directly or adventitiously because of these forces.

18.9.2 Familial Amyloidotic Polyneuropathy

Familial amyloidotic polyneuropathy (FAP) is an autosomal dominant disease characterized by a distal to proximal symmetrical polyneuropathy . This neuropathic pattern involves sensory and motor nerves in addition to a severe autonomic neuropathy. Most of the autosomal dominant amyloidoses are associated with mutations in plasma transthyretin (TTR). Transthyretin or prealbumin is a 127 amino acid residue single chain polypeptide that is made predominantly in the liver. Functionally, tetramers of transthyretin form rapidly following secretion from the cell. The tetrameric form of transthyretin is a transport protein that carries thyroxine and retinoic acid in association with retinoic binding protein. The monomeric TTR has extensive β-pleated sheet structure in antiparallel configuration. This secondary structure is thought to predispose the molecule to form the β-fibrillar structures characteristic of amyloid deposition. Approximately 54 mutations of the TTR primary sequence are known, most are associated with a painful peripheral and autonomic neuropathy with variable involvement of the heart, kidneys, gastrointestinal tract, and eyes. Symptoms of the disease usually begin in the fourth and fifth decade with relentless progression of the autonomic dysfunction, polyneuropathy, and occasional cardiomyopathy and renal failure until death commonly results within 10 years from malnutrition, complications of the autonomic dysfunction, or cardiomyopathy.

Significant advances in elucidating the chemistry underlying the aggregation of both wild-type and mutant TTR into amyloid fibrils have been made in recent years.

Many of these findings have contributed to the knowledge of the variety of amyloid diseases caused by the formation of β-fibrillar deposits in the tissues. Essentially, wild-type TTR (WT-TTR) in its tetrameric form does not form amyloid fibrils, but the WT-TTR, when treated with acid, undergoes a pH rate-dependent denaturation to several forms of monomer/dimer. One of these forms rapidly self-assembles into amyloid fibrils in a process whose rate of formation is also pH dependent. A fast equilibrium exists between the denaturation and production of amyloidogenic TTR monomers and the self-assembly process.

18.9.3 Spongiform Encephalopathies

The name of this group of amyloid-associated diseases is derived from the characteristic pathological finding in the brains of its victims: a degenerative process in which the cortical gray matter becomes riddled with vacuoles and suffers loss of neurons and proliferation of the supporting or glial cells. Under the microscope the brain looks like a sponge because of all of the vacuoles, hence the name spongiform. Amyloid plaques composed of an abnormal protein are found in all of these diseases whether the host is human or other animals. The spongiform encephalopathies are classified as subacute diseases because the clinical course tends to run over a period of several years before death ensues. However, the infectious agent thought to cause the clinical disease is usually acquired many years prior to the development of the clinical signs. These diseases are, therefore, considered slow infections. Humans suffer from three forms of this amyloidogenic syndrome. Four other forms have been described in animals.

The clinical pictures of kuru, CJD, and GSS are those of a moderately rapidly progressive dementia. Kuru is an infectious disease that formerly affected many children and adolescents in the isolated region of Papua New Guinea. The transmission was interrupted over three decades ago, and today the disease is found only among older adults in the area. This dramatic change in the epidemiology of kuru strongly supports the view that it was the practice of ritual cannibalism, especially of the brain, that was the most important if not the only mechanism of transmission of the disease among the people of New Guinea.

The earliest signs in kuru come from destruction of the cerebellum. The patient develops balance problems with unsteadiness which rapidly progresses to general incoordination. Typically the patient is seized with coarse shivering tremors. These tremors are the manifestation of the disease that gave kuru its name, derived from the Fore language. The degenerative process spreads variably to the adjacent brainstem, and patients lose their ability to control eye movements and to control the muscles of swallowing. Because of the inability to swallow and protect the airway most patients die from starvation or aspiration pneumonia. The severe incoordination often leads to accidental burns. The bedridden patients often develop decubitus ulcers which lead to septic shock and death. Patients usually die within 1 year from the start of clinical symptoms.

Creutzfeld–Jakob disease (CJD) similarly progresses to death within a year, often with signs of cerebellar involvement in some varying degree. Patients who develop the disease are on average in their sixties and present with disturbances in sensation, confusion, and inappropriate behavior. A progressive loss of mental faculties occurs over weeks to months, with the patient becoming comatose and thereafter death follows closely. In addition to the variable development of incoordination, most patients develop sudden jerking movements especially when startled, called "generalized startle myoclonus." Mean survival in CJD is 1 year, though 10% of patients are alive at 2 years. CJD occurs in three patterns, iatrogenic infection which is quite rare, in families, and as a sporadic disease. Gerstmann–Straussler syndrome is a familial form of CJD which tends to have a somewhat slower clinical course and involves more prominent cerebellar degenerative signs with a somewhat slower clinical course and a later dementia. Finally fatal familial insomnia is also an amyloid disease that runs in families (Table 18.9).

Table 18.9 Causes of the prion diseases

Disease	Etiology
Kuru	Infection
Creutzfeld–Jakob	
Iatrogenic	Infection
Sporadic	Unknown
Familial	PrP mutation
Gerstmann–Sträussler–Scheinker	PrP mutation
Fatal familial insomnia	PrP mutation

The clinical picture of CJD is very similar to that of scrapie which affects sheep and goats. Scrapie was so named because of the animals scraping their heads on fences when afflicted with the disease. Scrapie is passed between sheep by the ingestion of infected tissues. The question of inter-species transmission of the scrapie agent is of concern in the case of bovine spongiform encephalopathy (BSE), which is also more quaintly known as "mad cow disease." Transmission across the species barrier from sheep to cow is thought to have occurred because of the former practice in the United Kingdom of feeding offal of scrapie-infected sheep to domestic cattle, zoo animals, and possibly domestic cats. There was great concern in the mid-1990s about whether the amyloidogenic scrapie agent could make another species leap into humans who might consume BSE-infected beef. Concern that this actually poses a human health problem has now largely faded in the decade since the initial presentation of the problem.

Within the species barriers, however, CJD is clearly able to be transmitted between humans. All of the spongiform encephalopathies are transmissible to an appropriate host by inoculation of tissues from affected animals. The infectious agent passes through filters of 50 nm or more, and dilution studies indicate that very small inoculum of the agent will result in billions of infectious units per gram of tissue, each of these new particles identical in physical and biological property to the original inoculum. It is now believed that the particles are made up entirely of an infectious protein. To most analyses the “prion” agents appear to act as if they were proteins, but this has not been observed in any other agent with biological infectivity. Today confidence remains about the following statements:

(1) The prion agents contain no nucleic acid that codes for their progeny.
(2) The only known component of the prion is a modified protein (PrP^{Sc}) that is encoded by a cellular gene (the natural protein is called PrP^{C}).
(3) This major and possibly sole component of PrP^{Sc} is a pathogenic conformer of PrP^{C}.

Amyloid fibrils found in CJD brains are comprised of 27–30 kDa proteins called *prion protein 27–30 (PrP27–30)*. These peptides co-purify with the infectivity of the diseases. Whether they are themselves the infectious agent or are intimately linked with the particle causing infectivity, it is clear that the proteins are a product of the host cell itself. PrP27–30 is a 55 amino acid residue glycoprotein containing neuraminic acid and inositol that is formed by N-terminal trimming of a larger precursor protein PrP^{Sc}. The amyloid produced from PrP27–30 has substantial β-pleated sheet structure and is birefringent in polarized light after Congo red staining. PrP^{Sc} is a pathogenic isoform of the natural prion precursor protein PrP$_{C}$ which is found on the surface of normal cells. The function of PrP^{C} is unknown. The major difference between PrP^{C} and PrP^{Sc} is one of secondary structural conformation. PrP^{C} has substantial α-helix conformation (42%) and little β-pleated sheet structure (3%). PrP^{Sc} has a dramatically different secondary structure with 30% α-helix and 43% β-pleated sheet as measured by Fourier transform infrared spectroscopy. PrP27–30 is even higher in β-pleated sheet structure (54%) and lower in α-helix (21%). The link between natural PrP^{C} and pathogenic PrP^{Sc} appears to be a conversion from α-helix to β-sheet conformation that occurs via passage of the PrP^{C} through a thermally accessible intermediate, PrP*. The formation of PrP^{Sc} from PrP* is normally insignificant with most of the transition state population either returning to the PrP^{C} form or being degraded. However, when PrP* forms in the presence of PrP^{Sc} the formation of PrP^{Sc} is enhanced and accelerated. As more PrP^{Sc} is formed a positive feed-forward cycle is attained in which increasing amounts of PrP^{Sc} are generated from the natural PrP^{C}. PrP^{Sc} is pathogenic and probably causes neurodegeneration without the formation of amyloid deposits. The amyloid is formed from the PrP27–30 peptides, which are proteolytically processed from PrP^{Sc} whose β-sheet conformation predisposes its aggregation with the formation of the amyloid fibrils associated with the disease state. The amyloid itself does not participate in the key process which is the conformational conversion from PrP^{C} to PrP^{Sc}. In familial cases of the spongiform encephalopathies, mutations in the gene are proposed to lead to production of a PrP^{C} that is particularly susceptible to proteolytic processing from the α-helical precursor protein to the pathological β-sheet-containing fragment. Spontaneous mutations and the rare accumulation of sufficient PrP* to form enough PrP^{Sc} to begin the autocatalytic process are proposed mechanisms for the sporadic diseases. In infectious cases, the infectious agent is an inoculum of PrP^{Sc} material that begins the autocatalytic process.

The key observation that links the progression of the disease, its temporal pace, and the infectivity of the protein is the following: Once PrP^{Sc} is formed and begins to accumulate, the PrP^{Sc} itself is capable of catalyzing the conversion of natural PrP^{C} into β-sheet PrP^{Sc}. Thus the presence of the pathological peptide is capable of activating a process that by all appearances is one of self-propagation but is much more nearly akin to nucleation and propagation of self-aggregating structures such as the extracellular matrix (collagen for example) or a membrane. The rate control of such self-propagating assemblies depends on concentration of precursor (thus the linkage to the PRNP gene to incubation time), the environmental suitability for self-assembly (a general physical control mechanism), the mechanism and rate of conversion to intermediates capable of assembly, the rate of incorporation into the growing matrix, and the rate of disassembly. As we will discover in the coming chapters, the kinetic tools and potential energy surfaces that we will use to describe and understand such processes are fundamentally the same in fine detail throughout the physical state space and differ not in principle but in local curvatures in the state space which give biological and chemical definition and diversity.

In summary, in the case of familial amyloidosis, the amino acid mutations allow physically induced deaggregation of the normal tetramers into monomer/dimers that then interact incorrectly and form amyloid fibrils rather than functional TTR. It is likely that relatively small changes in tertiary and secondary structures make this path toward amyloid formation accessible. Similar proposals are accumulating evidence for similar misfolding pathways in Alzheimer's disease and light chain amyloidosis. Finally the conversion of the mutated prion protein between the natural α-helix form and the β-sheet form which is predisposed to form amyloid fibrils appears to provide the necessary potential energy template on which other normal proteins can be induced to misfold themselves. The key to understanding the folding and misfolding problem both in normal and pathological functions is an appreciation of the dynamic state of the protein and the shape of the potential energy surface over which the protein moves.

Further Reading

General

Flory P.J. (1953) *Principles of Polymer Chemistry.* Cornell University Press, Ithica, NY. (Remarkably this book is still considered by many to be the "bible" of polymer chemistry. It is still in press and available in 2009 without alterations since its writing over 50 years ago.)

Sun S.F. (1994) *Physical Chemistry of Macromolecules, Basic Principles and Issues.* Wiley, New York.

Tanford C. (1961) *Physical Chemistry of Macromolecules.* Wiley, New York.

Protein Folding

Creighton T.E. (ed.) (1992) *Protein Folding.* W.H. Freeman, New York.

Dill K.A. (1985) Theory for the folding and stability of globular proteins, *Biochemistry,* **24**: 1501–1509.

Elber R. and Karplus M. (1987) Multiple conformational states of proteins: a molecular dynamics analysis of myoglobin, *Science,* **253**:318–321.

Gruebele M. (1999) The fast protein folding problem. *Annu. Rev. Phys. Chem*., **50**:485–516.

Kaiser E.T. and Keidzy F.J. (1983) Secondary structures of proteins and peptides in amphiphilic environments (a review). *Proc. Natl. Acad. Sci.* USA., **80**:1137–1143.

Kim P.S. and Baldwin R.L. (1982) Specific intermediates in the folding reactions of small proteins and the mechanism of protein folding, *Annu. Rev. Biochem.*, **51**:459–489.

Matthew C.R. (1993) Pathways of protein folding, *Annu. Rev. Biochem.*, **62**:653–683.

Onuchic J.N., Luthey-Schulten Z., and Wolynes P.G. (1997) Theory of protein folding: the energy landscape perspective. *Annu. Rev. Phys. Chem*., **48**:545–600.

Richards F.M. (1991) The protein folding problem, *Sci. Am.*, **264, 1**:54–63.

Rose G.D. and Wolfenden R. (1993) Hydrogen bonding, hydrophobicity, packing and protein folding, *Annu. Rev. Biophys. Biomol. Struct.*, **22**:381–415.

Scholtz J.M. and Baldwin R.L. (1992) The mechanism of α-helix formation by peptides, *Annu. Rev. Biophys. Biomol. Struct.*, **21**:95–118.

Shea J.-E. and Brooks III C.L. (2001) From folding theories to folding proteins: a review and assessment of simulation studies of protein folding and unfolding. *Annu. Rev. Phys. Chem*., **52**:499–535.

Tarek M. and Tobias D.J. (2002) Role of protein-water hydrogen bond dynamics in the protein dynamical transition. *Phys. Rev. Lett.*, **88**:138101.

Taubes G. (1996) Misfolding the way to disease, *Science*, **271**:1493–1495.

Timasheff S.N. (1993) The control of protein stability and association by weak interactions with water: how do solvents affect these processes? *Annu. Rev. Biophys. Biomol. Struct*., **22**:67–97.

Udgaonkar J.B. (2008) Multiple routes and structural heterogeneity in protein folding, *Annu. Rev. Biophys.,* **37**:489–510.

White S.H. and Wimley W.C. (1999) Membrane protein folding and stability: physical principles. *Annu. Rev. Biophys. Biomol. Struct*., **28**:319–365.

The Amyloidoses

Benson M. and Wallace M. (1995) Amyloidosis, In Scriver C. et al. (eds.) *The Metabolic Basis of Disease*, McGraw-Hill, New York.

Bergethon P.R., Sabin T.D., Lewis D., Simms R.W., Cohen A.S., and Skinner M. (1996) Improvement in the polyneuropathy associated with *familial amyloid* polyneuropathy after liver transplantation, *Neurology,* **47**:944–951.

Caughey B., Baron G.S., Chesboro B., and Jeffrey M. (2009) Getting a grip on prions: oligomers, amyloids and pathological membrane interactions, *Annu. Rev. Biochem,* **78**:177–204.

Eigen M. (1993) Viral quasi-species, *Sci. Am.*, **269, 1**:42–49.

Ferreira S.T. and De Felice F.G. (2001) Protein dynamics, folding and misfolding: from basic physical chemistry to human conformational diseases. *FEBS Lett.*, **498**:129–134.

Kelly J.W. and Lansbury P.T. (1994) A chemical approach to elucidate the mechanism of transthyretin and *β* -protein amyloid fibril formation, *Amyloid*: *Int. J. Exp. Clin. Invest.*, **1**:186–205.

Kemper T.L. (1994) Neuroanatomical and neuropathological change during aging and dementia, In Albert M.L. and Knoefel J.E. (eds.) *Clinical Neurology of Aging*, 2nd edition. Oxford University Press, Oxford, pp. 3–67.

Prusiner S.B. and DeArmond S.J. (1995) Prion protein amyloid and neurodegeneration, *Amyloid: Int. J. Exp. Clin. Invest.*, **2**:39–65.

Selkoe D.J. (1991) Amyloid protein and Alzheimer's disease, *Sci. Am.*, **265, 5**:68–78.

Steinhart C. (1996) Sick cows, protein geometry and politics, *J. Chem. Educ.*, **73**:A232–A233.

Macromolecular Interaction and Cell Regulation

De Camilli P., Emr S.D., McPherson P.S., and Novick P. (1996) Phosphoinositides as regulators in membrane traffic, *Science*, **271**:1533–1539.

Fuchs E. and Weber K. (1994) Intermediate filaments, *Annu. Rev. Biochem.*, **63**:345–382.

Görlich D. and Mattaj I.W. (1996) Nucleocytoplasmic transport, *Science*, **271**:1513–1518.

Grunstein M. (1992) Histones as regulators of genes, *Sci. Am.*, **267,4**:68–74.

Luscombe N.M., Laskowski R.A., and Thornton J.M. (2000) An overview of the structures of protein-NA complexes, *Genome Biol.*, **1**:1–37.

Luscombe N.M., Laskowski R.A., and Thornton J.M. (2001) Amino acid-base interactions: a three dimensional analysis of protein-DNA interactions at the atomic level, *Nucleic Acids Res.*, **29**:2860–2874.

Luscombe N.M., Laskowski R.A., and Thornton J.M. (2002) Protein-DNA interactions: amino acid conservation and the effects of mutations on binding specificity, *J. Mol. Biol.*, **320**:991–1009.

Rothman J.E. and Orci L. (1996) Budding vesicles in living cells, *Sci. Am.*, **274, 3**:70–75.

Schatz G. and Dobberstein B. (1996) Common principles of protein translocation across membranes, *Science*, **271**:1519–1526.

Scherman R. and Orci L. (1996) Coat proteins and vesicle budding, *Science*, **271**:1526–15322.

Vallee R.B. and Sheetz M.P. (1996) Targeting of motor proteins, *Science*, **271**:1539–1544.

Problem Sets

1. Consider the concept that the primary structure leads to a functional protein as described by Anfinsen's experiments with ribonuclease. Using an elementary catastrophe cusp show why Anfinsen's experiment could work with ribonuclease but would not have been so "predictable" if insulin had been used.
2. Calculate the solvent molar and volume fractions as well as the polymer molar and volume fractions for each of the following polymer solutions given that the solvent molecular size is 0.3 nm and monomer molecular size is 0.3 nm:

	Size of polymer	Polymer:solvent ratio
a.	1000mer	1:1000 polymer:solvent ratio
b.	100,000mer	1: 10 polymer:solvent ratio
c.	10mer	1:1000
d.	5mer	1:1000
e.	monomer	1:10

3. Assume that the polymer segments are distinguishable. How many configurations could each of the polymer chains in Problem 2 form?

4. Calculate the configuration number for each of the polymer chains in Problem 2 if the polymer segments are indistinguishable. Comment on the reasons for the difference.
5. Show that differentiating $\Delta G_{\mathrm{mix}} = kT\left[\chi n_{\mathrm{s}}\phi_{\mathrm{p}} + \left(n_{\mathrm{s}} \ln \phi_{\mathrm{s}} - n_{\mathrm{p}} \ln \phi_{\mathrm{p}}\right)\right]$ with respect to n_{s} is the same as $\Delta u_{\mathrm{solvent}} = -RTV_{\mathrm{s}}^{\mathrm{o}}\left(\frac{1}{FW_{\mathrm{p}}}c_p + \frac{V_{\mathrm{s}}^{\mathrm{o}}}{2FW_{\mathrm{p}}^2}c_{\mathrm{p}}^2 + \cdots\right)$.

Chapter 19
Molecular Modeling – Mapping Biochemical State Space

Contents

19.1 The Prediction of Macromolecular Structure and Function Is a Goal of Molecular Modeling

At this point in our journey we should be in general agreement about the central element of biophysical chemical study: the important biological observables of function and action in a biological state space are a direct consequence of the coordinate structure of the physical elements (mass, energy, and forces) of biomolecules. We have used this concept to construct potential energy surfaces which connect the position of the physical elements in space with a measurable interaction, i.e., force

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_19, © Springer Science+Business Media, LLC 2010

or energy. How this helps us with our interest in function is as follows. The observed function of a system is simply the perceived interaction of the system with elements within the system and with the observer (whether strongly or weakly coupled to the system). All interactions require energy. If no force or action is exerted between elements of a system the system can have no function. Furthermore, if no force is exerted by the system on the observer it is impossible to assign function to the system.

Given this perspective it is natural to look for its practical application at a structural level. If we know the forces and energies acting between simple systems such as atoms and molecules we should be able to construct a potential energy diagram of these forces and energies. Then we can relate energy to position and predict the structure of a new larger system composed of smaller elements and their interactional linkages. Theoretically, we should be able to predict new structures and once the form of these structures is at hand we should be able to predict their properties and understand their function. Since most of the interactional energies are relatively straightforward calculations we would expect that a fairly detailed picture of function could be drawn from basic physical principles. It is this linkage between chemical physics and biology that holds the appeal and promise of molecular modeling as an invaluable tool for the biophysical chemist.

19.2 Molecular Modeling Is Built on Familiar Principles

Molecular modeling is a predictive tool that represents the movement in the modeling mirror diagram (Fig. 4.1) from right to left (i.e., from the formal system toward a prediction of the natural system). Our ability to make accurate predictions from first principles or ab initio may be limited by a set of abstractions and their related states that do not reflect the full set of linkages and observables necessary to describe the natural system properly. However, in the case of molecular modeling and computational chemistry this is probably not the problem. The principles of quantum electrodynamics are probably sufficiently detailed and dependable that we could QED to provide strikingly accurate real predictions from first principles. The problem is that the calculations are so difficult that we cannot implement our knowledge in any practical fashion. Likewise Dirac’s relativistic quantum mechanics is not implementation friendly and most commonly we use the Schrödinger equation as our quantum abstraction. We have already discussed this equation in some detail including the Born–Oppenheimer approximation which allows us to create a potential energy surface for a given set of nuclear coordinates that is an adequately practical approximation for most chemical purposes. The Schrödinger equation provides the foundation recipe for most calculations of molecular structure that attempt to remain consistent with quantum mechanical assumptions. As we already appreciate, exact analytical solutions to the Schrödinger equation are practically impossible and all of the quantum mechanical methods require simplifying assumptions of one sort or another. The practical game is to choose the correct degree of simplification for the given problem at hand without generating an unacceptable degree of inaccuracy.

The potential energy surfaces that are sought in molecular modeling experiments are essentially mechanical in nature, that is, they relate the actions and related energies of the system being modeled to positions in a multidimensional coordinate space. Therefore at times it is more convenient to use various subsets of quantum mechanical treatments including Newtonian mechanics as the "rules of engagement" for a particular problem. This separation into quantum and Newtonian mechanical approaches makes a logical separation of molecular modeling techniques into ab initio and *empirical methods*. The most popular ab initio methods are all based on the molecular orbital methods of LCAO-MO that have been described in a previous chapter. These methods are completely quantum mechanical in their approach and even though a wide variety of simplifications may be applied in an attempt to solve the integrals needed in the LCAO-MO method, they are ab initio because they solve the integrals by direct solution rather than using experimental or empirical data to simplify the solution of the integrals. In some cases empirical data are used and hybrid methods which are also drawn from a quantum mechanical origin are called *semi-empirical methods*. The ab initio and semi-empirical methods actually do not make any a priori assumption of such things as chemical bonds but rather the bonding behavior is a result of the quantum mechanical calculations.

The empirical methods which include ball and stick models as well as molecular force field methods utilize the ideas of bonding derived from these quantum mechanical calculations but assign specific properties to the bonds linking atoms and the atomic properties of the objects linked in coordinate space based wholly on measured physical properties. Empirical methods are built from the body of physical measurements such as x-ray crystallography, calorimetry, and infrared spectroscopy. These techniques provide essentially mechanical properties of molecules which are used in a Newtonian framework to model molecules of interest.

Today elaborate visually compelling displays of molecules are common and it is difficult to appreciate how modern the ideas of atomic and molecular bonding are. The concept of valence and its implied bonding which underlies our modern understanding of chemistry and biochemistry is less than 200 years old. We are familiar with representations of molecular structure that include empirical (C_2H_5OH) and structural formulas and already have used them extensively in this volume. Now we go beyond these representations and explore the empirical, force field, semi-empirical, and ab initio methods of molecular modeling in use today.

19.3 Empirical Methods Use Carefully Constructed Physical Models

19.3.1 Sticks and Stones

Two primary modes of physical models are useful when building molecular structures: *crystallographic* and *space-filling models*. In crystallographic models the bond lengths, bond angles, and the crystallographic radii of atoms are used to build a precise model of the covalent framework that connects atoms into molecules. The most popular of these is probably the ball and stick model in which the backbone structure of a molecule can be well appreciated by visual inspection. While the ball and stick abstraction suffers because it ignores the actual space occupied by the atoms in the molecule while emphasizing the covalent skeleton, the Dreiding model is an abstraction in which only the covalent structure is demonstrated completely at the expense of the topographic information. We often wish to go beyond the covalent backbone of a molecule and develop a sense of what the molecule must look like to an observer of the same dimension as the molecule. In other words what is the actual physical space taken up by the molecule. Space-filling models are an attempt to provide this information. When a cell, enzyme, or receptor interacts with another molecule, it feels the physical dimension of volume because it is excluded from that volume. In like fashion the enzyme or other "observer" molecule sees the electric field, charge distribution, and fluctuations of the electronic structure that give each element of the biomolecule its chemical nature. Images of these models are shown in Fig. 19.1.

19.3.1.1 Study Question

These pictures are pretty and very compelling. As we proceed we will notice that the renderings will become even fancier and more compelling. This is partly because we have no practical experience seeing molecules and so it is very difficult to perform a reality check on the overpowering stimulus of the visualizations. Before we become too enamored, an exercise with these very simple models is in order.

Consider this simple question: If we are drawing models based on "accurate" knowledge of the radii of atoms, the bond distance between the atoms, the angles of those bond distances, and the freedom of rotations of those bond distances, how comfortable are we with the method(s) of measurement? Which method gives the answer for the following?

a. Atomic radii
b. van der Waals radii
c. Interatomic distance
d. Interatomic bond angles
e. Bond character (single, double, triple)

Answers: (a) x-ray crystallography; (b) calculation of the energy minima from an expression like those in Chapter 9; (c) x-ray crystallography, spectroscopy; (d) x-ray crystallography; (e) spectroscopy.

From this exercise, it is obvious that the observables necessary to construct a ball and stick model are largely available from relatively direct measurement. The task of building a space-filling model requires knowledge of the interaction volume for the elements of the biomolecule. This is different from the electronic volume which can be measured by the x-ray interaction with the electron clouds of atoms.

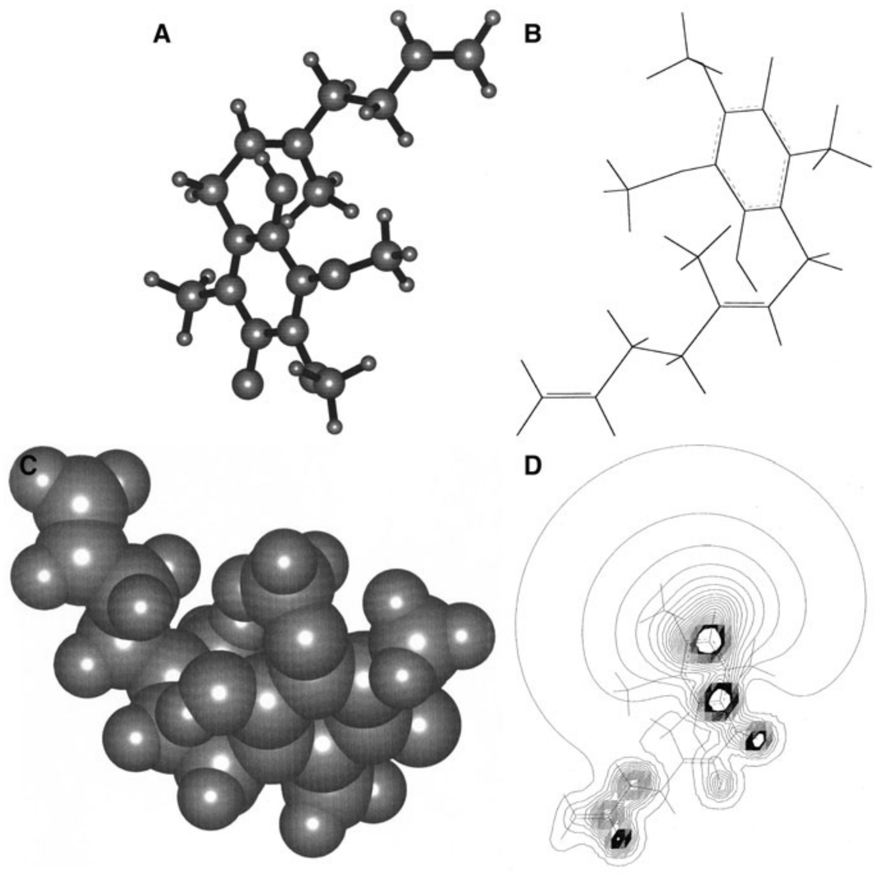


Fig. 19.1 Examples of various "ball and stick"-type molecular models of the semiquinone form of the electron transport cofactor ubiquinone. (**a**) Ball and stick; (**b**) Dreiding (framework); (**c**) Space filling; (**d**) This is a framework model with a quantum mechanical calculation of the electrostatic potential superimposed on the molecular structure

19.3.2 The Ramachandran Plot Is the "Art of the Possible"

The geometries of molecules are related to their potential energy and the relationship can be plotted on a potential energy surface. A reasonable generalization is that molecules will move toward the lowest overall potential energy and this allows us to pick a preferred geometrical conformation from a potential energy plot. A simple illustration is shown in Fig. 19.2. We compare the rotation around a central carbon–carbon *sp*^{3}-single bond for ethane, CH_3CH_3, and *n*-butane, $CH_3CH_2CH_2CH_3$. Since an *sp*^{3} bond is cylindrically symmetrical there can be no energy difference between the various positions taken by the terminal methyl or ethyl groups, yet the potential energy surface clearly demonstrates that different energies are associated with each

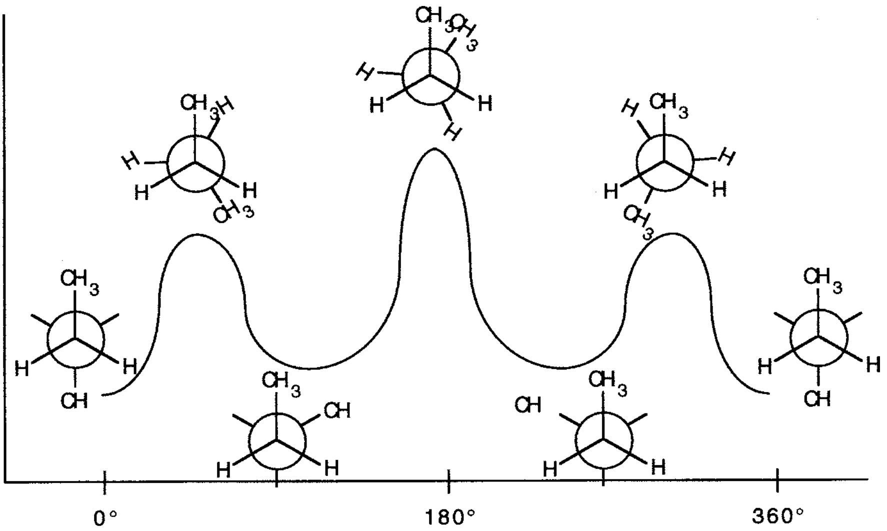


Fig. 19.2 Simple model to demonstrate potential energy (*y* axis) of steric interactions in *n*-butane. There are three maxima and three minima. The *anti* configuration is of lower energy than the *gauche* conformation

of the possible conformations. In this pair of cases the energy over the observed minima is due to the steric interference as the electronic distributions of the hydrogens in the case of ethane and the methyl groups in the case of *n*-butane which move in and out of contact with one another. A similar potential energy surface can be described for any coordinate trajectory including torsional strain because of rotation around a bond with π electron distribution or out-of-plane bending in which the atomic coordinates are moved away from the equilibrium molecular geometries. The role played by a complete analysis of this type will be explored shortly when we examine the methods of molecular mechanics. Now we consider only the limitations imposed by steric hindrance and by the permitted rotation around bonds with σ and π nature.

We know that with formation of peptidyl bond linking amino acids into a polypeptide chain, every third bond in the backbone has some double bond character and is not free to rotate. Free rotation of a pair of restricted, planar C–N peptide bonds around the central or α-carbon can theoretically occur because of the cylindrically symmetrical nature of the σ bonds connecting the α-carbon to the pair of amide-bonded groups. We perform an analysis of the steric interactions experienced by the polypeptide. Reading from the N-terminal to C-terminal end of the polypeptide the two planar elements will rotate around; the N–C$_a$ σ bond whose angle of rotation is measured as ϕ, and the C$_\alpha$–C σ bond whose angle of rotation is measured as ψ. The details of these measurements are shown in Fig. 19.3. In the extended form of a polypeptide $\phi = \psi = 180°$. If ψ is fixed at +180° and ϕ rotates to approach 0°, further rotation is prevented because of the steric hindrance

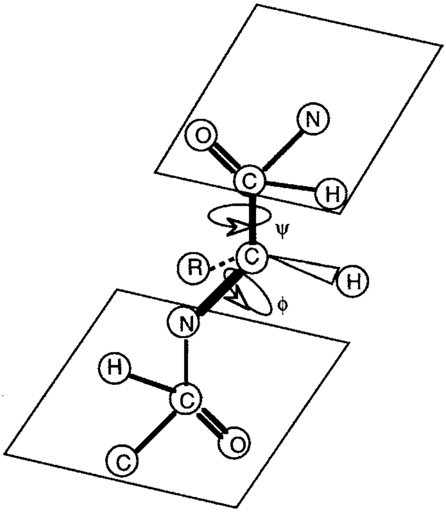


Fig. 19.3 Definition of the ϕ and ψ angles for conformational analysis of peptide interactions. A ball-and-stick model of two peptide bonds connected to a central atom that connects to an *R* group and a hydrogen belonging to the right-handed residue. The plane of the two peptide bonds lies in the plane of the page making the structure represented the open or extended form. The *R* group extends into the plane of the page and the H sticks out from the page. The ϕ and ψ angles may each independently take any value from –180° to 180°. The easiest way to remember how to assign values to ϕ and ψ is as follows. Using the right-hand rule grab the σ bond under consideration such that your thumb points away from the C_α. The values of ϕ and ψ are the angle subtended when the planar peptide is rotated in the direction in which your fingers are curled around the bond. By convention the angles in the extended form are $\phi = \psi = +180$ deg. The extended form can be recognized by the appearance of the anagram $\begin{matrix} \text{H O} \\ \text{O H} \end{matrix}$

between the carbonyl carbons. Alternatively when ϕ is fixed at +180° and ψ rotates to approach 0°, further rotation is prevented because of the steric hindrance between the hydrogens of the peptidyl nitrogens. When ϕ=180° and ψ = +120°, the C_α–H bond is *trans* to the C=O bond. When ϕ = 120° and ψ = +180°, the C_α–R bond is *trans* to the (N–H) bond. This analysis demonstrates that each pair of peptide bonds will have constraints on the freedom of rotation around the cylindrically symmetrical σ bonds of the C_α. There will be angles that will be prohibited and certain angles that will give energetically preferred conformations. By taking these constraints into consideration, G.N. Ramachandran devised a plotting system in which a steric contour diagram or Ramachandran plot representing the two observables, ϕ and ψ, could be used to exclude the energetically unlikely conformations. These graphs plot ϕ on the abscissa and ψ on the ordinate. A single plot is made for each amino acid residue because there will be interaction between the R groups of each C_α thus giving each residue its own steric contour plot (Fig. 19.4). As would be expected, there will be forbidden sets of angles as well as preferred and intermediate zones of permitted conformation. These zones of permitted structure can be

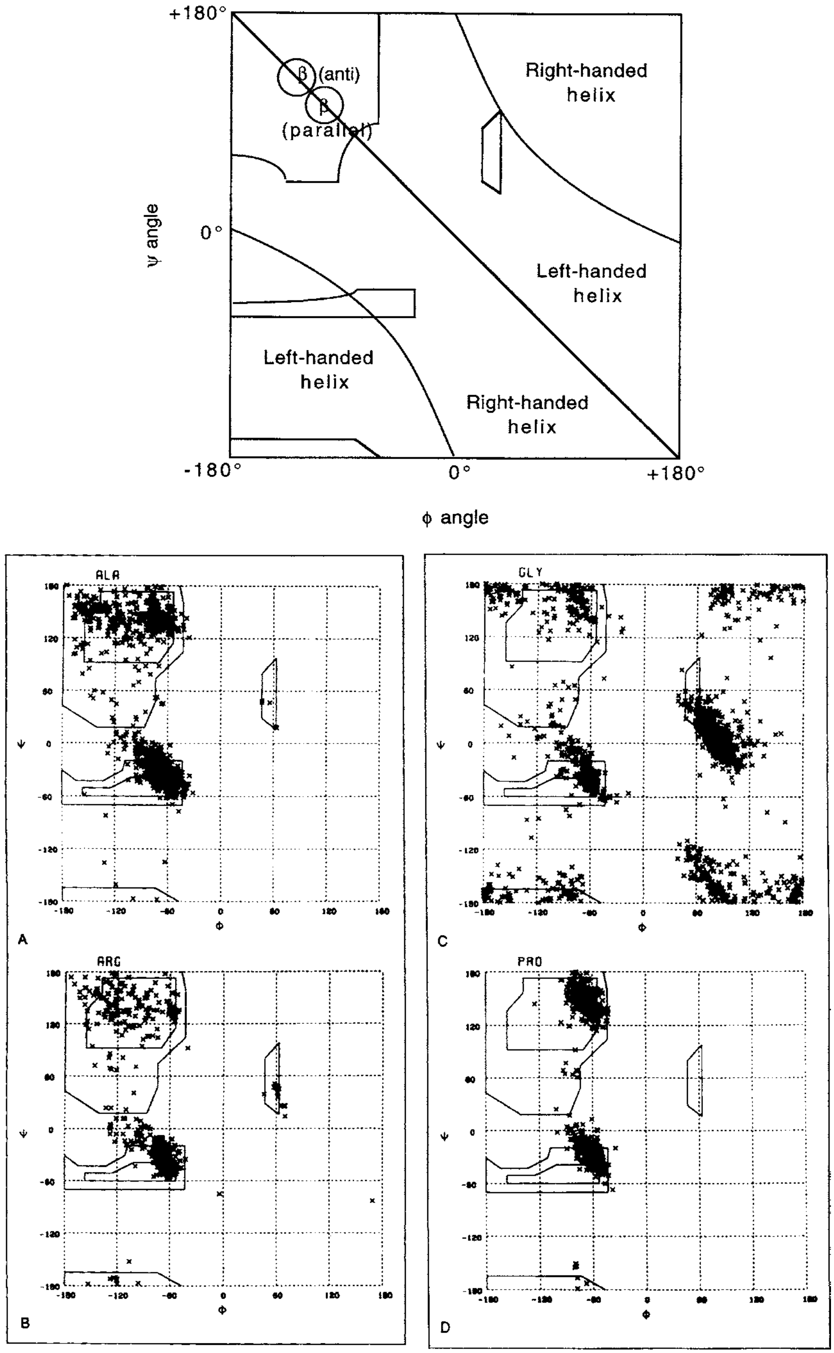


Fig. 19.4 Ramachandran plot showing angles associated with certain secondary structures; **(b)** plots by computational methods for glycine, arginine, proline, and alanine (From Finzel B.C. et al. (1990) in Bogg C.E. and Ealick S.E. *Crystallographic and Modeling Methods in Molecular Design*. Springer-Verlag, Berlin.)

Table 19.1 Compilation of torsion angles for polypeptide structures

Structure	ϕ	ψ
α-Helix, right handed [α-poly-(L-alanine)]	–57°	–47°
α-Helix, left handed [α-poly-(L-alanine)]	+57°	+47°
3_{10}helix	–20	–54
β-Pleated sheet, parallel	–119°	+113
β-Pleated sheet, antiparallel	–139°	+135°
Polyglycine	–80	+150
Polyproline II	–78	+149

From IUPAC-IUB Commission on Biochemical Nomenclature, *Biochemistry*, 9:3471, 1970

compared with compilations of the torsional angles for regular polypeptide structures (Table 19.1) and thus be used to predict whether a given residue is likely to be found in a particular type of secondary structure.

The data to construct the Ramachandran plots come from contact radii for various atomic pairs taken from x-ray crystallographic analysis of various polypeptides (Table 19.2). By constructing an accurate model in terms of the bond lengths and angles (given in Table 19.3) and using these contact radii, we can

Table 19.2 Contact distances for polypeptide atomic pairs

Pair	Normal (nm)	Outer Limit (nm)
C–C	0.32	0.3
C–O	0.28	0.27
C–N	0.29	0.28
C–H	0.24	0.22
O–O	0.28	0.27
O–N	0.27	0.26
O–H	0.24	0.22
N–N	0.27	0.26
N–H	0.24	0.22
H–H	0.2	0.19

Table 19.3 Geometry of the polypeptide chain

Bond	Length (nm)	Bond angle
C_α–C_ψ	0.153	–
C_ψ–N	0.132	–
N–C_α	0.147	–
C_ψ=O	0.124	–
N–H	0.100	–
C_α–C_β	0.154	–
C_α–H_α	0.107	–
C_α–C_ψ–N	–	113°
C_ψ–N–C_α	–	123°
N–C_α–C_ψ	–	110°

demonstrate that the plot for glycine is accurate. Given the geometry and trigonometry of the peptide bond it should be obvious why a different plot exists for each amino acid. Take the case of alanine. The allowable conformations are substantially reduced compared with glycine because there are many more steric constraints when the R group is CH_3 as in alanine compared to the restrictions found when $R = H$.

The steric contour diagram is able to predict that the permitted conformations account for all known regular structures of naturally occurring polypeptides. Recently similar analysis has been performed using computer modeling and quantum mechanical treatments. Except for minor differences, the general contours of the two generations of plot are the same. In spite of its power, the steric contour plot is a very simple model limited to local short-range phenomena. We can easily make it fail and the demonstration is valuable. The steric contour plots for alanine, glutamic acid, and lysine all show that an α-helical structure is permitted. However, we now perform an experiment using polypeptides made from multiple repeating residues of each of these, i.e., poly-L-alanine, poly-L-glutamic acid, and poly-L-lysine. When poly-L-alanine is added to an aqueous solution of pH 7, it spontaneously forms α-helices. Yet when a similar solution of poly-L-glutamic acid or poly-L-lysine is made, these polyamino acids are found in a random chain arrangement with no α-helix formation. Investigation of the steric contour plot provides no information regarding this "surprising" result. Remember that surprises occur when a description of a state space is incomplete. The equation of state (in this case predicting the formation of the α-helix) bifurcates with a dependence on an unmeasured observable. Since the steric contour plots explore short-range interactions in state space perhaps the first step should be to consider longer range interactions that might affect helix formation. Choosing the simplest and longest range interaction first, the electrostatic force, we note that because of their ionizable groups, both of these polyamino acids will contain large amounts of like charge at pH 7. It is a reasonable hypothesis that this long-range interaction that is not accounted for in a steric contour plot is contributing to the "surprising" observation that the α-helix is not formed. We can test this hypothesis by an experiment in which we remove the charge on the polyamino acid chains by adjusting the pH. Noting that the pK_a of the glutamic acid side chain is 4.25 and that of lysine amino acid chain is 10.53. We choose a pH of 2 and 12, respectively, to effectively remove the charges on the respective polyamino acid chains. At these conditions, both polyamino acids form α-helices spontaneously.

Thus, the steric contour plots that limit the possible conformations of a polypeptide based only on short-distance peptidyl interactions are of value but are not complete. The chemical nature of the side chain residues in addition to their space taking properties is important in the overall structure. These properties must be represented on the potential energy surface for biomolecules. We have already alluded to some of these properties and shown how they can be used to classify the amino acid residues. Let us see how taking these properties into account in the empirical model-making process can be useful.

19.3.3 Secondary Structure Prediction in Proteins Is an Important Challenge in Molecular Modeling

As we have now seen the physical interactions of ion–ion, van der Waals, and hydrogen bonding interactions (which are responsible for the stabilization of the helix and sheet structures) can be approached based on empirical measurements. A fundamental difference between empirical methods and ab initio methods of model building is that ab initio methods require no adjustment or accounting for the chemical environment since the environment is also treated with first principles. Empirical methods use parameters that are taken from experiment and may be environment specific thus making parameter sharing between modeling exercises risky. Nevertheless, for biological systems this empirical information is often the only information available and it is widely used. The responsibility for its appropriate and prudent use inevitably falls onto the individual user and researcher.

An area of importance in computational chemistry is the full prediction of the secondary and higher structure of a polypeptide. This is a fundamental goal of theoretical molecular biology and is an area of active research. In general one starts with a pair of observations:

(1) Native proteins in an aqueous environment are organized so that their charged and polar side chains are exposed to the solvent (water) and hence are found on the exterior of the protein.
(2) There are two secondary structure motifs, the α-helices and the β-sheets, which are formed in order to satisfy the condition of maximal hydrogen bonding.

These two constraints are generally met in all native proteins. However, one needs to appreciate that if internal hydrogen bonding of the amino and carbonyl moieties of the peptidyl backbone are insufficient, there will not be adequate free energy to ensure the protection of the hydrophobic side chains from the solvent. As a consequence of these two simultaneously operating requirements, we would expect to find that most proteins are composed of segments of helix and β-sheet linked together by relatively flexible linkers or loops. These loops will allow maximal packing of the secondary structure hence the tertiary structure. From topological considerations we would neither expect the secondary structure segments nor the loops to be too long, a condition that would prevent effective overall folding. An examination of the solved x-ray crystallographic structures indicates that the great majority of known natural protein structures fit this description.

The secondary structural elements in proteins are minimally stable as helices or sheets; if an isolated small section of peptide of specific secondary structure is removed from its overall protein environment it will frequently denature. Thus the overall protein structure is a state in which the properties of the system are dependent on the system itself and not just from the addition of each of the individual elements. Practically there is cooperativity between the backbone hydrogen bonding and the overall hydrophobic effect in the global folding of a protein. This system-wide property makes the modeling of protein folding very challenging. Ultimately, the goal is to describe the potential energy surface whose contour leads to a global minimum potential energy well that represents the most likely protein structure. There are three overall strategies for computing protein structure:

(1) Atomic-level energy calculation for a protein based on the physiochemical interactional energies. We have discussed the foundations of this approach and will explore its application in force field and quantum mechanical models shortly. This approach is direct and intellectually appealing but is computationally expensive for very large molecules.

(2) The prediction of protein structure as a consequence of the amino acid sequence is based on empirically derived rules. These rules are derived from a statistical analysis of a database containing the structural information about known native protein structures. Using various properties of the amino acids such as their hydrophobicity or patterns of secondary structure interaction in tertiary state space, these methods make predictions based on an equation of state that relates observables drawn from a very specific state space.

(3) A hybrid modeling system lying between these models specifies the spatial organization of the peptide chain based on empirical data from a database. It then approximates the side chain and intermolecular interactions using physiochemical and empirical data in a search for the overall tertiary folding path.

We will discuss aspects of the first and second of these strategies.

The simplest of the empirical methods (strategy 2) asks whether a particular amino acid is found on the protein surface in an aqueous solvent or if it is found buried away from solvent. Thus the major properties considered are the hydrophobicity and hydrophilicity of the residue. This is then coupled with observations derived from a database in which a particular amino acid is associated with a secondary structure as a (1) former, (2) breaker, or (3) indifferent participant in that structure. The successful use of this type of analysis is associated with the work of Chou and Fasman but has been used by others as well. Assignment of amino acids to these structures based on analysis of a specific set of proteins is shown in Tables 19.4 and 19.5. The predictive accuracy of these methods is at best in the range of 70%. The basic model can be extended to include secondary structure interactions drawn from x-ray crystallography and NMR data. Using a set of proteins to build a predictive database, these model systems attempt to restrict the abstract state space in which a model may be located by considering how the overall structure stabilizes the local secondary structures. Based on our earlier discussion this extension has substantial theoretical appeal.

We conclude with a description of the scales used in empirical models to predict secondary structure. All of these scales reflect to some degree the hydrophobicity/hydrophilicity of the amino acids. The scales we will consider are

Table 19.4 Amino acids as formers, breakers, and indifferent participants in different types of secondary structure

	α-Helix tendency		β-Sheet tendency	
Residue	P_α	α-Assignment	P_β	β-Assignment
Ala	1.45	H_α	0.97	I_β
Arg(+)	0.79	i_α	0.90	i_β
Asn	0.73	b_α	0.65	b_β
Asp(–)	0.98	i_α	0.80	i_β
Cys	0.77	i_α	1.30	h_β
Gln	1.17	h_α	1.23	h_β
Glu (–)	1.53	H_α	0.26	B_β
Gly	0.53	B_α	0.81	i_β
His (+)	1.24	h_α	0.71	b_β
Ile	1.00	I_α	1.60	H_β
Leu	1.34	H_α	1.22	h_β
Lys(+)	1.07	I_α	0.74	b_β
Met	1.20	h_α	1.67	H_β
Phe	1.12	h_α	1.28	h_β
Pro	0.59	B_α	0.62	b_β
Ser	0.79	i_α	0.72	b_β
Thr	0.82	i_α	1.20	h_β
Trp	1.14	h_α	1.19	h_β
Tyr	0.61	b_α	1.29	h_β
Val	1.14	h_α	1.65	H_β

H_x = strong former; h_x = former; I_x = weak former; i_x = indifferent; b_x = breaker; B_x = strong breaker. $P_x = f_x/<f_x>$ where f_x is the frequency of residues in the helix and β-regions and $<f_x>$ are the average frequency of residues in the regions for 15 proteins analyzed
Data from Chou and Fasman, *Adv. Enzymol*. (1978) 47:45.

(1) *Hydropathy* – reflects the tendency for a residue to be found in the interior versus the exterior of a protein in an aqueous environment. Hydropathy is based on the partition coefficients of an amino acid between pure water and an organic phase. The free energy of transfer between the two phases can be calculated from the partition coefficients. A dielectric constant appropriate to the protein interior is chosen for the calculations. Different assumptions lead to a choice of dielectric constant between that of hexane (3) and ethanol (20). Examples of a hydropathy plots are shown in Fig. 19.5.

(2) *Antigenicity* – reflects the likelihood that a particular amino acid will be an antigenic determinant. Since antibodies recognize surface features of macromolecules this scale reflects the predilection for an amino acid to find its way into the solvent interfacial region. These scales are hydrophilicity scales based on polar–apolar solvent partitioning since it is the chemical nature of the charged hydrophilic side chains that leads them to have the biological activity of immunogenicity. Essentially this scale is a reciprocal scale of the hydropathy plot. The linkage between structure and biological function is simply an inverse

Table 19.5 Synopsis of the Chou–Fasman rules for secondary structure prediction

I. Classify and assign each residue according to Table 19.4

II. Search for helical regions

III. Search for β-sheet regions

IV. Search for overlapping α- and β-regions

V. Apply the following rules:

1. *Helical regions are predicted* when a segment exists with

a) ≥six residues with $\langle P_\alpha \rangle \geq 1.03$ and $\langle P_\alpha \rangle > \langle P_\beta \rangle$ and

b) II (a) through II (e) are satisfied

2. *β-sheets are predicted* when a segment exists with

a) ≥three residues with $\langle P_\beta \rangle \geq 1.05$ and $\langle P_\beta \rangle > \langle P_\alpha \rangle$ and

b) III (a) through III (e) are satisfied

3. Overlaps can occur with these rules and can be solved by a relatively complex algorithm found in the listed reference

II *Search for helical regions*

a. *Nucleation*

1. 4 out of 6 h_α or H_α residues initiate a helix

2. $I_\alpha = \frac{1}{2}\, h_\alpha$

b. *Propagation*

1. Extend the helical segment in both directions as long as the condition that the next tetrapeptide selected is not a terminator sequence (see (c))

2. Overlapping nucleation segments are linked together into a long helix

3. $I_\alpha = \frac{1}{2}\, h_\alpha$

4. Nucleated helix segments should have $> \frac{2}{3}$ helix formers

5. Propagating helix segments should have $> \frac{1}{2}$ helix formers

6. Nucleation + propagation segments contain $< \frac{1}{3}$ helix breakers

c. *Termination*

1. A propagated helix is terminated by tetrapeptide breakers if $\langle P_\alpha \rangle < 1.00$: $b_4, b_3i, b_3h, b_2i_2, b_2ih, b_2h_2, bi_3, bi_2h, i_4$

2. An adjacent β-region with $\langle P_\beta \rangle > \langle P_\alpha \rangle$ also terminates the helix

d. *Helix breaker*

1. Proline cannot be found in the inner helix

2. Proline cannot be found at the C-terminal end

3. Proline may be the 3rd residue in the N-terminal end (first turn)

e. *Changes at helix boundaries*

1. N-terminal: Pro, Asp(–), Glu(–) are incorporated in the helix. Pro, Asp(–) are given I_α assignments

2. C-terminal: His(+), Lys(+), and Arg(+) are incorporated in the helix. Arg(+) is given an I_α assignment

III *Search for β-sheet regions*

a. *Nucleation*

1. 3 β-formers in a row initiates a sheet

2. A cluster of 3 β-formers out of a segment of 4 or 5 initiates a sheet

b. *Propagation*

1. Extend the sheet in both directions as long as the condition that the next tetrapeptide is not a breaker sequence (see (d))

2. Overlapping nucleation segments are linked together into a long sheet

3. Propagating sheet segments should have $> \frac{1}{2}$ sheet formers

6. Formation is unlikely with $\geq \frac{1}{3}$ sheet breakers

Table 19.5 (continued)

c. *Termination*
1. A propagated sheet is terminated by tetrapeptide breakers if $\langle P_\beta \rangle < 1.00$:$b_4$, b_3i, b_3h, b_2i_2, b_2ih, b_2h_2, bi_3, bi_2h, i_4
2. An adjacent α-region with $\langle P_\alpha \rangle > \langle P_\beta \rangle$ also terminates the sheet

d. β-*sheet breaker*
1. Glu and Pro are strong sheet breakers. They should occur in a sheet only when they are an element of a tetrapeptide with $\langle P_\alpha \rangle < \langle P_\beta \rangle >1$

e. β-*sheet boundaries*
1. All charged residues and Pro are unfavorable to β-sheet formation. They occur in sheet only when they are an element of a tetrapeptide with $\langle P_\alpha \rangle < \langle P_\beta \rangle >1$

function to a measured physiochemical property. Such plots can be altered if database information relating to empirically known antigenic determinants is included.

(3) *Solvent accessibility* – reflects the hydrophilicity of an amino acid. Scales of this type are structurally empirical and are calculated from the atomic coordinates of a set of globular proteins. For a given amino acid, a ratio is determined:

$$\frac{\text{\# of residues} > 95\% \text{ buried}}{\text{total \# of residue}}$$

This number is inversely correlated to the hydropathy and some models use scales of this type to empirically correct the hydropathy scale.

(4) *Side chain mutability* – reflects the frequency with which a residue is changed in a set of homologous proteins throughout the evolutionary tree. Based on the principle that a residue at the surface is less likely to disrupt the native folding of a protein, this scale assumes that the highly conserved amino acids are found in the interior of the protein and the mutated ones are on the surface.

(5) *Backbone flexibility* – is based on thermodynamic factors that reflect the atomic motion in crystal structures. A scale of this type looks at a window of the primary sequence and scales a central amino acid by weighting the flexibility of the local structure on either side of the central moiety. A property is assigned based on whether it has 0, 1, or 2 “rigid” neighbors. This rigid neighbors are usually considered to be: ALA, LEU, HIS, TYR, ILE, PHE, CYS, TRP, or MET.

Scales like these can then be applied to a primary sequence and used to predict properties of segments of amino acids within the primary sequence. For example, the secondary structure and antigenicity for sperm whale myoglobin are compared for the experimental and predictive investigations in Fig. 19.6 and Table 19.6. Figure 19.6 demonstrates an important caution with the use of predictions of secondary structure. The crystallographic structure is compared to structures predicted by a training set of proteins. It can be appreciated that the choice of training set can strongly influence the secondary structure prediction often to an unreasonable degree. Furthermore even though secondary prediction may be reasonable, the tertiary structure is not at all reasonably predicted in this approach.

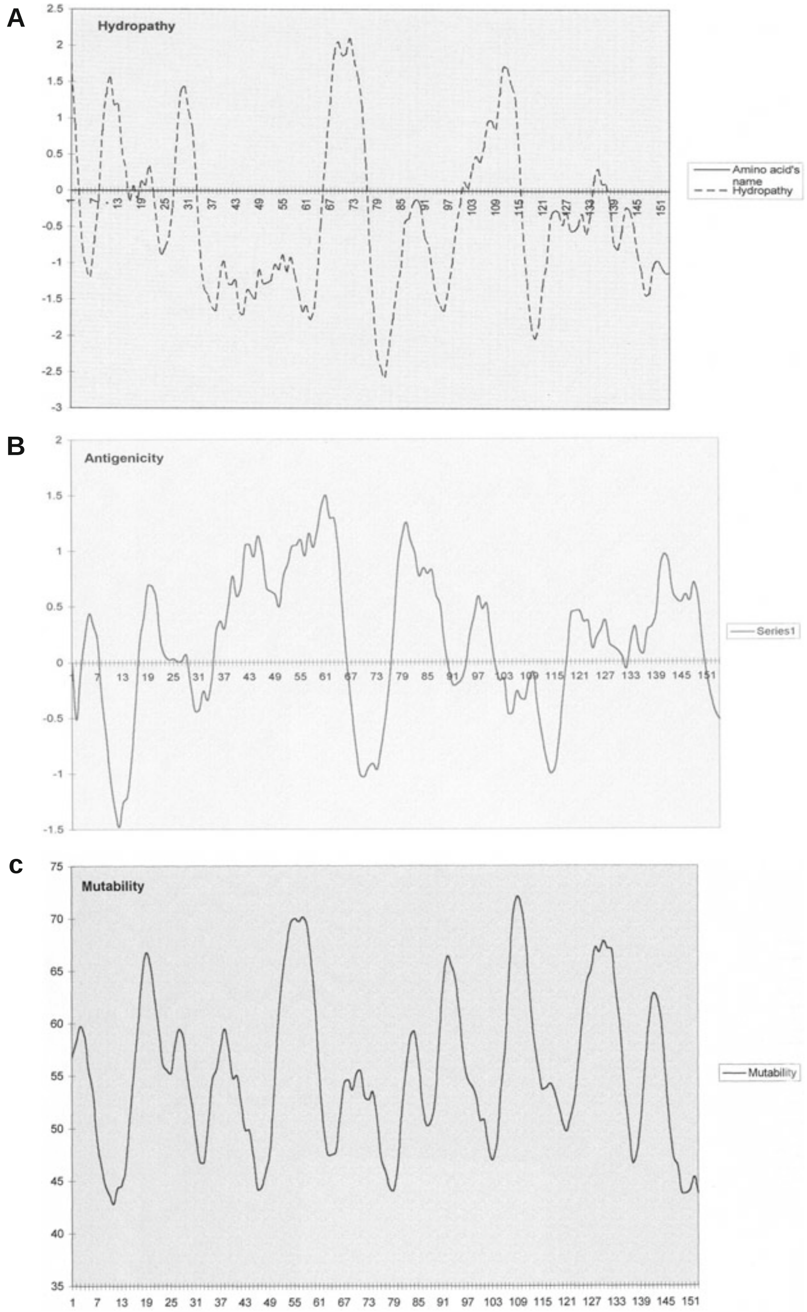


Fig. 19.5 Association plots of sperm whale myoglobin showing: (**a**) hydropathy; (**b**) antigenicity; (**c**) side chain mutability; (**d**) flexibility; (**e**) solvent accessibility

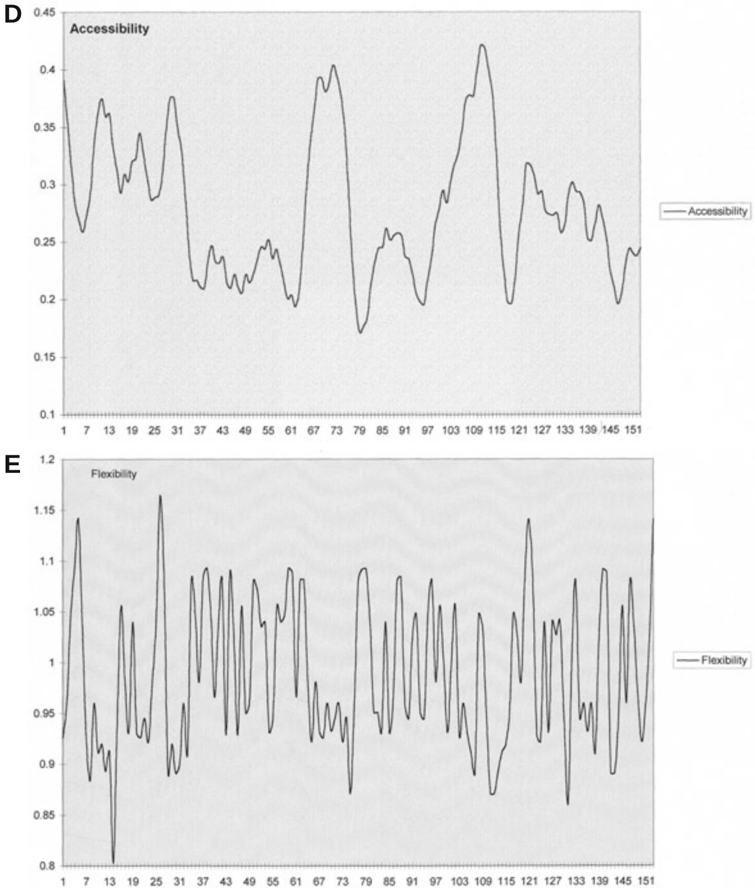


Fig. 19.5 (continued)

19.4 Computational Methods Are the Ultimate Gedanken Experiments

We have described the relative positions of atoms in a biomolecule in terms of restricted, permitted, and likely contour maps. These are derived from steric considerations which are the short-range interactional energies of volume exclusion; a component of the Lennard–Jones potential. The Ramachandran or steric contour plot is a potential energy plot in which only one interactional energy is considered as the various torsional angles of each peptide pair are considered. We can just as easily consider all of the other interactional energies in terms of a rotation around the ϕ and ψ angles which yields an equation of the following form:

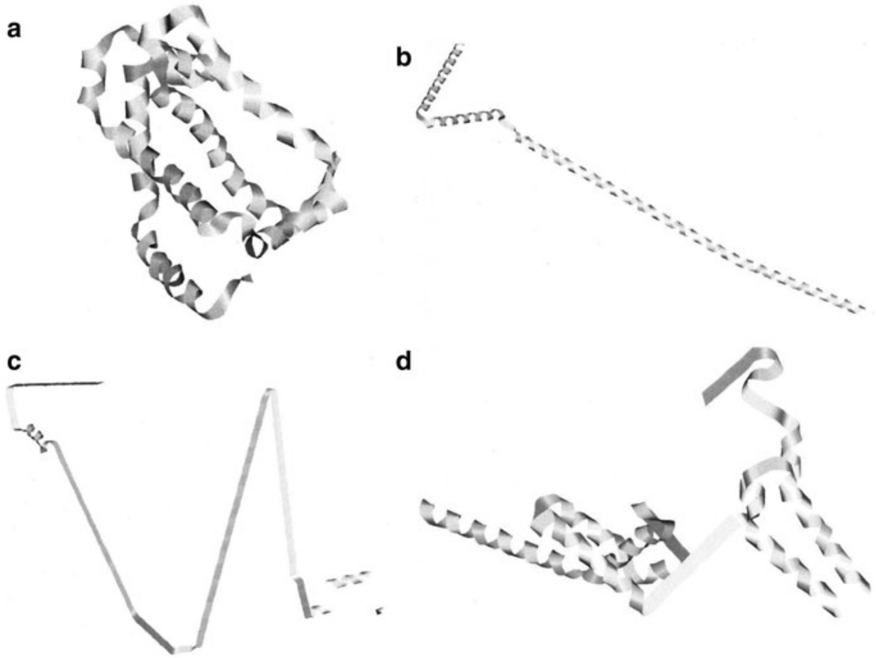


Fig. 19.6 Predicted secondary structure compared to native structure by x-ray crystallography for sperm whale myoglobin. In (**a**) the structure seen in x-ray crystallographic studies; in (**b–d**) various training sets are used to predict the likely secondary structure of the protein and the overall tertiary structure is generated randomly as the secondary structure data are read into the visualization program. (**b**) α-helix training set yields dominantly α-helix; (**c**) β-sheet training set is heavily weighted toward β-sheets; (**d**) a mixed training set provides reasonably accurate secondary structure prediction but has no tertiary structure predictive power

Table 19.6 Comparison of predicted (from Fig. 24.7) versus experimental antigenicity for sperm whale myoglobin

Regions of antigenicity	
Experimental	Predicted
# of residues from C-terminal	# of residues from C-terminal
15–22	18
56–62	60
94–99	96
113–119	118
145–151	147

Data from Hopp T.P. and Wood K.R. (1981)*PNAS*, USA, 78:3824.

$$E\left(\phi_i,\psi_i\right)=\sum\left(E_{vdW}\left(\phi_i,\psi_i\right)+E_{i-i}\left(\phi_i,\psi_i\right)+E_{d-d}\left(\phi_i,\psi_i\right)+\cdots\right) \tag{19.1}$$

We can generalize the above discussion to include the construction of potential energy surfaces for molecules of substantial complexity. These methods use either quantum mechanical and classical mechanical approaches and enable us to construct a potential energy surface by determining the forces and energies associated with the different arrangements of atoms. These techniques are built from the expressions of interactional energy that we have developed in the previous chapters and are valuable because of the relatively abundant and inexpensive computational power available today. In general the computational techniques of quantum and molecular mechanical modeling have one purpose: to determine the energies and the energy derivatives (forces) necessary to produce a potential energy surface that can then be used for a wide variety of purposes. Some of these applications include

(1) Visualization and display of molecular geometry and structure,
(2) Determination of optimal structures of molecules,
(3) Visualization of orbital wave functions, electronic distributions, and properties that derive from these quantum mechanical treatments,
(4) Determination of thermodynamic and molecular properties of molecules,
(5) Investigation of reactivity of molecules,
(6) Evaluation of possible kinetic and mechanistic paths of chemical reactions,
(7) Study of dynamic behaviors including ligand binding, enzyme activity, and conformational changes during reactions.

Generally all molecular and quantum mechanical methods use the Born–Oppenheimer approximation. Energies are generated on the assumption that the nuclear coordinates are fixed for the purposes of the calculations.

19.5 Molecular Mechanics Is a Newtonian or Classical Mechanical Modeling Approach

Molecular mechanical methods use classical analytical functions in their treatment of the atoms and their interactions to construct a molecular potential energy surface. The equation that is used to calculate the potential energy for each possible arrangement of atoms in the coordinate space providing a description of the molecule is called a *force field*. There are a variety of force fields in current use each using somewhat different mechanical functions and different interaction parameters. Therefore different force fields may be more appropriate for varying systems. Yet virtually all force fields determine the potential energy of a molecule based on the positions of the atoms and their distance dependent energies. Thus the potential energy of a molecule will be calculated from the sum of interactional energy terms:

$$E_{\text{total}} = E_{\text{b-st}} + E_{\text{ang}} + E_{\text{tor}} + E_{\text{vdw}} + E_{\text{oop}} + E_{\text{el}} + E_{\text{Hbond}} + E_{\text{di-di}} \qquad (19.2)$$

In force field methods, the total energy calculated has no real physical meaning but rather is a difference term compiled by comparison of the calculated partial energy of each term with the parameterized minimum or equilibrium term for, respectively, bond stretching, bond angle, torsional strain around a bond, van der Waals interactions (usually both attractive and repulsive or steric), out-of-plane bending, electrostatic, hydrogen bonding, and dipole–dipole interactions. While a single force field calculation has little intrinsic physical value it is useful as a comparison between molecules. Though there is a relationship between the enthalpy of a molecule and a force field energy calculation, there is no consideration of the thermal motion and the temperature-dependent contributions to the total energy. It is therefore impossible for E_{total} to be an enthalpic term. Force field calculations are a specialized abstraction in the sense that they treat all atoms and bonds as balls connected by springs according to some force–distance function. There is no explicit consideration of the electrons and so phenomena that are dependent on electronic distribution such as bond formation and breaking, molecular orbital interactions, and electronic delocalization will not be seen.

Given these constraints why use force field methods? The obvious reason is that computationally it is relatively easy to solve a many parameter equation such as Eq. (19.2) using classical approximations and this allows treatment of very large molecules such as proteins and nucleic acids. The molecular qualities are imparted to the method by experimentally derived values for the force constants and in many cases this results in computed values for molecules that are closer to experiment than are the values computed with the theoretically more complete and accurate ab initio quantum mechanical models. In practical experience, force field methods are often used in conjunction with other empirical methods (such as the semi-empirical quantum mechanical methods) and ab initio methods to solve computational problems. Force field models do not generalize well since the molecular parameters are developed for a specific class of molecules in a specific environment. They may be extremely accurate within those constraints but the force field will often fail if applied without care to systems outside those constraints.

The practical use of a molecular mechanics model requires three interacting components: (1) the function, (2) the atom type and, (3) the parameter set. The first of these is the *mathematical* or *functional form* of the force field which is composed of the specific functions comprising Eq. (19.2). These contain important assumptions and approximations which we will explore shortly. The force field function acts on atoms. But, how do we practically describe the atomic population? Since all of the interactions will be described by analytical terms like those in Eq. (19.3), every pair of atoms could have their own specific force constants and equilibrium geometry. Though an ideal description of the chemical state space would contain this detail, such a solution is impractical. Furthermore any application of the force field to a generalized or unknown case is lost. Thus an abstraction is applied in order to maintain both generality and practicality. Instead of specific atom-by-atom descriptions, atoms are assigned an *atom type* from a restricted list. For example, carbon would be distinguished as *sp*, sp^2, sp^3, or aromatic, and the force constants and bond geometries would be assigned between atoms of each type. Interactions are then calculated not between individual atoms in a molecule but rather between the types of atoms in the molecule. Atom types can be defined in terms of any property that distinguishes atoms based on their chemical environment though the most commonly used properties are probably hybridization, immediate bonded neighbors, and formal charge on the atom.

The concept of atom types is central to the development of molecular mechanical methods. In some cases it is practical to treat a collection of atoms as an atom type. This step is usually taken such that the hydrogens are bonded to carbon atoms. This is a computational economy compared to the case in which each bond is considered explicitly. The carbon–hydrogen group can be defined as an atom type with the carbon–hydrogen bonds treated implicitly. A force field utilizing this technique is called a *united force field* and treats methyl, methylene, and methine groups. The increased size of the C–H grouping will be reflected in increased van der Waals radii in the parameter sets. Force fields that treat each hydrogen explicitly are called *all atom force fields.*

Having applied the categorical abstraction of the atom type to describe the molecular system, a set of parameters must be assigned to the various constants called for by the functional form of the force field. These variables comprise the *parameter set* of a force field model and are usually derived from experimental data such as spectroscopic measurements. In some cases the numbers in a parameter set may be found by ab initio computational methods and then used in a force field calculation. Parameter sets are generally accurate for a specific class of molecule measured under a constrained set of experimental conditions. New molecules may be modeled with some confidence using a particular parameter set if they belong to the class for which the parameter set was developed and are being evaluated under similar environmental constraints. For example, the MM2 force field of Allinger and co-workers is parameterized for small organic molecules while the AMBER force field developed by the Kollman research group is designed for work with proteins and nucleic acids. A summary of some of the force fields in current use is given in Table 19.7.

Table 19.7 Comparison between force fields

Force field	Originators	Unified/all atom	Parameter sets
MM2	Allinger	All atom	Small organics
AMBER	Kollman	Both	Proteins and nucleic acids
CHARMM	Karplus	Both	Macromolecules
OPLS	Jorgensen	Unified atom	Proteins and nucleic acids

Table 19.8 Interaction types included in the interactional potential

Bonded	
Direct	1–2 bond stretching
Geminal	1–3 bond bending
Vicinal	1–4 dihedral angle rotation
Non-bonded	
Exchange repulsion	Pauli exclusion of electron distributions (repulsive interaction of the van der Waals terms)
Dispersion attraction	Attractive portion of the van der Waals term
Electrostatic	Charge, dipole, and quadrapole interactions
Cross terms	
Interactional terms between the above forces such as bond angle–bond stretch terms	

Now we will look at the functional terms of the force field equation in more detail. The interaction potential includes both bonding and non-bonding interactions. The types of these interactions are summarized in Table 19.8.

19.5.1 Bond Stretching

The abstraction used in force field modeling is that of atoms connected together by springs. The simplest stretching motion of such a system is a harmonic function whose potential energy is given by Hooke's law:

$$U_{\text{stretch}} = \sum_{\text{bond}} K_r (r - r_0)^2 \tag{19.3}$$

The problem with this function as an abstraction is apparent if Fig. 19.7a is considered. The harmonic function is a symmetrical function that suggests that the potential energy of two atoms approaching one another is identical to the energy of two atoms exceeding the bond length and thus separating from one another. We know that when two atoms approach closely to one another their interactional potential energy increases at a rate much steeper than that given by the harmonic function. In addition once the bond length exceeds the equilibrium bond distance, the energy is much flatter than described by the harmonic function. Furthermore at an internuclear distance somewhat less than twice the equilibrium bond length, the bond no longer exists and the potential becomes that of a dissociating bond. These features are present in Fig. 19.7b in the rendering of the *Morse potential*. The Morse potential is a function that has many of the characteristics of the bond potential that we have just discussed, but is much more computationally expensive to use in a force field system. It is clear that the harmonic function adequately represents the bond stretching potential only at internuclear distances close to the equilibrium length. Often polynomial functions of higher order than the quadratics are added to

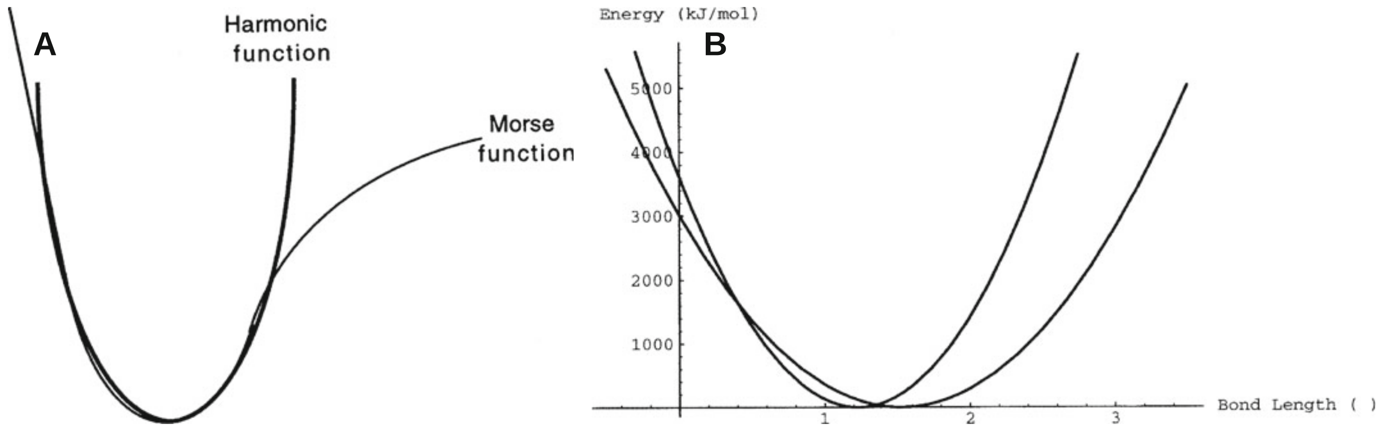


Fig. 19.7 The harmonic function as a model for the bond stretching energy. (**a**) The harmonic function generates a symmetric potential energy curve that only approximates a portion of the more accurate Morse potential curve. (**b**) The effect on the shape of the harmonic function caused by varying K_r can be seen in this illustration of a carbonyl C–O bond and an aliphatic C–C bond taken from the parameter set of the CHARMM force field

improve the fit of the stretching term to the more ideal Morse potential. An important practical application of knowing the type of stretching function used in a force field application must be emphasized. A force field computation that uses a harmonic function must be started from a structure that is close to its equilibrium bond lengths or else substantial unrealistic forces can be assigned to the off-equilibrium atoms.

19.5.2 Bond Bending

For small displacements from the equilibrium bond angle, the bond bending potential term is often treated with a harmonic function since the bending may be thought of like a coil spring being bent onto itself:

$$U_{\text{bond}\angle} = \sum_{\text{angles}} K_\theta \left(\theta - \theta_0\right)^2 \tag{19.4}$$

As the bending force constant term K_θ increases, the tendency of the system to remain at the equilibrium bond angle θ_0 increases. Though a function such as the Morse potential does not exist for bond angle bending often higher exponential terms such as cubics and quartics can be added to improve the description of the bending energy.

19.5.3 Torsional or Dihedral Potential Functions

We saw earlier that bonded interactions between 1 and 4 atoms often have preferred conformations such as the *gauche*, *cis*, and *trans* conformers. These preferred conformations are modeled by rotating the dihedral angles through a 360° path. A truncated Fourier series is often used to generate the dihedral potential function.

$$U_{\text{dihedral}} = \sum_{\text{dihedral}} \begin{bmatrix} \frac{V_1}{2}\left(1+\cos\left(\psi-\phi_0\right)\right) + \frac{V_2}{2}\left(1-\cos\left(2\psi-\phi_0\right)\right) \\ +\frac{V_3}{2}\left(1+\cos\left(3\psi-\phi_E\right)\right) + \frac{V_n}{2}\left(1+\cos\left(n\psi-\phi_0\right)\right) \end{bmatrix} \tag{19.5}$$

Here V_{n} is the dihedral force constant, n is the periodicity of the Fourier term, ϕ_0 is the phase angle, and ψ is the dihedral angle. Wavefunctions of arbitrary complexity can be generated by adjusting the force constants, the phase angles, and the periodicity. Many force fields are truncated at two or three periodicity terms. As an example consider the amide portion of the peptidyl system's torsion (HN–C=O) as calculated by the AMBER and force field (Fig. 19.8). The description in the AMBER force field includes Fourier periodicities of 1 and 2 with a phase shift of 0 and 180°. The potential curve with $n = 1$ and $\psi = 0$ has a maximum at 0° and minima at ±180°. This potential curve has $n = 2$ and $\psi = 180$ and has two maxima one each at ±90° and three minima, at 0 and ±180°. The amplitudes of the force

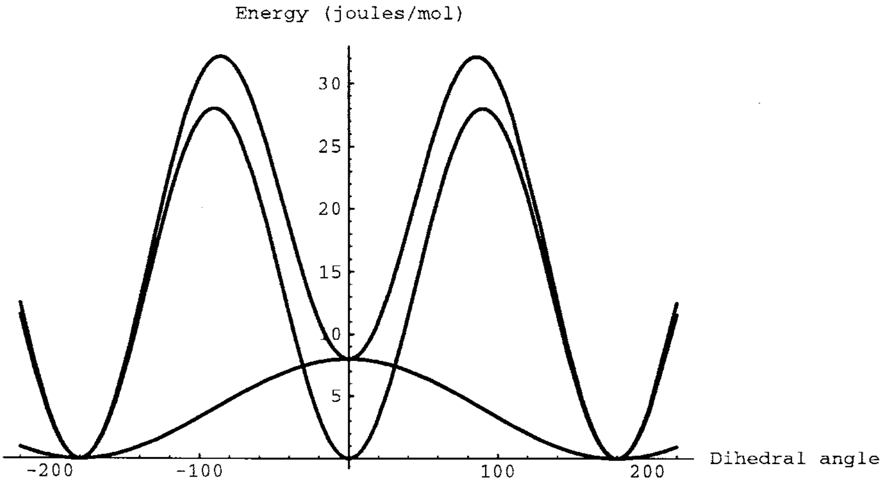


Fig. 19.8 Dihedral angle potential energy function for the amide bond (reproduced from Computational Chemistry, Hypercube, Inc. (1993) with permission)

constants V_1 and V_2 represent the energy barriers between the rotamer states. When summed the two curves give a resultant which represents the rotation through the dihedral angle where an angle of 0° represents a *cis* state and angle of 180° is the *trans* configuration. Note that the potential energy curve shows that the energy of the *cis* conformer is higher, thus less stable than the *trans* conformer. More terms can be added to the function to refine the available conformer states.

19.5.4 van der Waals Interactions

In many force fields the Lennard–Jones potential as described in Chapter 9 is used to describe the short-range repulsive and long-range attractive components that represent the van der Waals interactions between non-bonded atoms.

$$U_{\text{interaction}} = \sum_{ij} \frac{B_{ij}}{r^{12}} - \frac{A_{ij}}{r^6} \tag{19.6}$$

The use of the 6–12 potential in force fields is an example of a compromise between a computationally convenient function (the 6–12) and a more physically accurate function such as the Buckingham potential that treats the repulsive portion of the interaction as an exponential function:

$$U_{\text{Buckingham}} = Ae^{(B/r)} - \frac{C}{r^6} \tag{19.7}$$

The MMx force fields of Allinger use the Hill equation which like the Buckingham equation uses an exponential function for the repulsive force. The interaction parameters that define the constants such as A and B in the force fields define either the minimum energy separation and the depth of the well for each paired or single atom type or alternatively A and B are derived from the radii of the atoms and a measure of “hardness” of the atom. The “hardness” is related to the steepness of the repulsive portion of the energy well. Further discussion of these terms can be explored in the references cited at the end of the chapter.

Between appropriate atom types, the hydrogen bond is frequently treated in a similar fashion to Eq. (19.6) except rather than using a 6–12 potential, the hydrogen bonding interaction is written as a 10–12 interaction:

$$U_{HB\text{interaction}} = \sum_{ij} \frac{C_{ij}}{r^{12}} - \frac{D_{ij}}{r^{10}} \tag{19.8}$$

Typically the potential energy well of a 10–12 function is deeper than that of a 6–12 potential with a somewhat closer approach distance for the hydrogen bonded pair which is the principle contribution of the interaction potential.

19.5.5 Electrostatic Interactions

The interactional energy of non-bonded electrostatic interactions is generally given by the classical treatment:

$$U_{\text{electrostatic}} = \sum_{ij} \left[\frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon r_{ij}} \right] \tag{19.9}$$

Generally this function is applied to monopole–monopole interactions between non-bonded atom types greater than a 1–5 relationship apart. 1–4 non-bonded interactions are often treated a little differently because they represent a boundary condition. The “boundary” treatment of 1–4 atom pairs in general applies to all of the interactions that we have discussed to this point. Other electrostatic terms like those we have already explored in Chapter 9 including dipole–dipole and even induced dipole interactions are often added to the electrostatic force field. The application of these functions is straightforward but an interesting and important question is brought to light by these functions that are not so straightforward. In a molecular system, how should the dielectric constant be treated. The dielectric constant is a property of the environment in which charges find themselves and usually represents the physical nature of a continuum environment. When we consider a molecular system in vacuo, a fixed value of the dielectric constant ($\varepsilon = 1$) is correct. However, systems of biological interest include solvent. We can still use the in vacuo value for ε if we explicitly include each water or other solvent molecule in our molecular modeling calculation but practically this is a difficult computational solution. As our discussion of the Kirkwood equation in Chapter 15 showed the continuum value of ε depends on the local structure of the solvent especially when the solvent is highly structured as in the case of water. This means that the dielectric constant of a solvent such as water is different in monolayer versus in the interphase versus regions with bulk properties. Generally in the case just described, ε: monolayer < interphase < bulk. Many force fields deal with this issue by using a distance-dependent dielectric to represent this solvent property implicitly.

The empirical force field is a recipe for approximating the energy of a molecule that assumes that the behavior of the molecule behaves according to the functional terms of the force field and that adequate parameterization is available for the atoms in the molecule. Force fields are used in three principle types of modeling programs. In molecular mechanics applications selected points on the potential energy surface of a molecule are approximated. Force fields can be used in *Monte Carlo* simulations where a variety of starting conditions are randomly chosen in order to study a distribution on a potential energy surface or in *molecular dynamics* simulations where the future positions and vectors of a molecular structure are found from the present positions. Such a simulation is accomplished by putting a molecular structure into motion on its potential energy surface.

19.6 Quantum Mechanical Methods Are Computational Difficult but Theoretically “Pure”

While molecular mechanical methods depend on the concept of atom types and the associated parameters, semi-empirical quantum mechanical calculations use parameters associated with only one atom or one atomic number. Given this relatively small set of parameters, the wavefunction for a molecule is calculated using quantum mechanical methods. Ab initio calculations determine the molecular structure strictly from the Schrödinger equation and the physical constants e, m_e, and $\hbar$ and and the atomic numbers of the constituent atoms. In both cases the Schrödinger equation is solved for the electronic Hamiltonian that results from applying the Born–Oppenheimer approximation to a molecular configuration. An approximate potential energy surface can be determined by repeated solutions of this type at a variety of nuclear configurations. Such a process which requires the calculation of the Schrödinger at each point can be very computationally intense and is very time consuming. This is particularly true for ab initio calculations on large molecular systems where detailed calculations are required. The semi-empirical methods ease the computational constraint somewhat by replacing some of the more difficult calculations with experimentally derived values. We will not explore the methods of computational quantum chemistry in this volume except to briefly list the generational evolution of the semi-empirical systems (see Table 19.9).

Table 19.9 The semi-empirical time line

Theory	Approximation	Originator
Hückel	Ignore σ interactions with π electrons, ignore overlap integrals, all remainingintegrals are expressed as parameters α and β; not a self-consistent field method	E. Hückel
Extended Hückel	Recognizes σ interactions with π electrons; otherwise uses similar approximations to above	R. Hoffman
NDO methods	*Neglect of differential overlap* All of the following methods ignore *differential* overlap rather than the overlap *integral,* this eliminates many electron–electron repulsion integrals	
CNDO	Complete Neglect of Differential Overlap	Pople
INDO	Intermediate Neglect of Differential Overlap	Pople
MINDO/3	Modified (INDO)	Dewar
NDDO methods	Neglect of Diatomic Differential Overlap	
MNDO	Modified Neglect of Diatomic Overlap	Dewar
AM1	Austin Method 1	Dewar
PM3	Reparameterization of AM1	Stewart

In the Chapter 8 we discussed the general approach to solving the Schrödinger equation for a molecular system. Semi-empirical methodologies apply these approaches and have several resulting features:

- The parameters used for each element are independent of the chemical environment in contrast to the dependence on the environment of parameters used in molecular mechanics systems. Usually the parameters in quantum mechanical methods are generated from experimental data or from ab initio calculations.
- These methods require no information about the organization or geometry of the bonds in a system. Practically, however, a quantum mechanical method can be speeded significantly if a reasonable structure is taken as the starting point. For this reason many problems start with a molecular mechanics minimization followed by a quantum mechanical method to calculate the wavefunction.
- Because semi-empirical methods directly provide information about electronic density and the probability pattern of a wavefunction, they can describe bond breaking.
- The wavefunctions that are generated from a quantum mechanical calculation contain the necessary information to calculate dipole moments, atomic charges, charge and spin density as well as the molecular orbitals of the molecule.

The availability of this information from a series of quantum calculations allows the potential energy surface to be explored for kinetic paths complete with prediction of transition states, bond breaking, and formation and molecular structure. In addition, quantum calculations make it possible to predict both thermodynamic and vibration, UV–visible, and EPR spectra.

Further Reading

General

HyperChem® computational chemistry: *Molecular Visualization and Simulation*. Hypercube, Inc. Publication HC40-00-03-00, 1994. (Guide tutorial to one of the widespread popular PC based all-in-one programs. Gives a generally useful review of the field of computational chemistry.)

Jensen F. (2007) *Introduction to Computational Chemistry*, 2nd edition. Wiley, Hoboken, NJ.

Lipkowitz K.B. and Boyd D.B. (eds.) (1990) *Reviews in Computational Chemistry.* VCH Publishers, New York. (A set of review articles covering the field of computational chemistry. A good place to start further investigation of the field. There are annual volumes with this name.)

Force Fields

Berkert U. and Allinger N.L. (1982) *Molecular Mechanics.* American Chemical Society Monograph 177, Washington, DC. (The classic introduction to force field methods.)

Boyd D.B. and Lipkowitz K.B. (1982) Molecular mechanics, the method and its underlying philosophy, *J. Chem. Educ.*, **59**:269–274.

Lipkowitz K.B. (1995) Abuse of molecular mechanics: pitfalls to avoid, *J. Chem. Educ.*, **72**: 1070–1075.

McCammon J.A. and Harvey S.C. (1987) *Dynamics of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge. (The application of force field calculations applied to biological macromolecules.)

Quantum Mechanical Methods

Hinchcliffe A. (1988) *Computational Quantum Chemistry.* Wiley, New York.

Dynamical Modeling

Daggett V. and Levitt M. (1993) Realistic simulations of native-protein dynamics in solution and beyond, *Annu. Rev. Biophys. Biomol. Struct.*, **22**:353–380.

Karplus, M. and Petsko A. (1990) Molecular dynamics simulations in biology, *Nature*, **347**: 631–639.

McCammon J.A. and Karplus M. (1980) Simulation of protein dynamics, *Annu. Rev. Phys. Chem.*, **31**:29–45.

Secondary Structure Prediction

Chou P.Y. and Fasman G.D. (1978) Prediction of the secondary structure of proteins from their amino acid sequence, *Adv. Enzymol.*, **47**:45–148.

Friesner R.A. and Gunn J.R. (1996) Computational studies of protein folding, *Annu. Rev. Biophys. Biomol. Struct.*, **25**:315–342.

Gilbert R.J. (1992) Protein structure prediction from predicted residue properties utilizing a digital encoding algorithm, *J. Mol. Graphics*, **10**:112–119.

Problem Sets

1. Is further modification of the Ramachandran plot dependent on the overall volume of the R group? Why or why not?
2. Propose an atom-type classification for oxygen that would be useful in biochemical investigations.
3. Compare the shapes of potential energy curves for a 6–12 versus a 10–12 function. Calculate and plot these functions using Eq. (19.8).

Chapter 20
The Electrified Interphase

Contents

20.1 The Interphase Is Formed When Phases Meet

Whereas homogeneous systems are relatively easily described, cellular processes are heterogeneous and more difficult to describe. Because so many processes occur that require the exchange of components across at least one phase, it is extremely valuable for the biological scientist to have an understanding of the forces and structures that act in the zone of transition between phases. When different phases come in contact with each other, an interface between them occurs. This interface is a surface, and the properties of a surface are different from those of either of the phases responsible for creating it. Additionally, the changeover between phases is never instantaneously abrupt, but instead there is a zone of transition extending from the surface for a finite distance into the bulk of each of the phases where the properties are representative of neither bulk phase. The surface and the regions immediately adjacent are termed the *interphase*, a very useful distinction. The properties of interphases will be the subsequent focus of this section.

The development of interphase regions is a thoroughly general phenomenon and applies throughout nature. The approach here will be to develop a qualitative picture

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_20, © Springer Science+Business Media, LLC 2010

of these interface regions in general and then to relate this picture to cellular systems. Much of the seminal work in understanding the nature of the interphase has been done in aqueous electrolyte – mercury drop studies, and, though not directly convertible, there is much valuable information that can be gained from considering this knowledge. Although the cell membrane is probably better considered as an insulator or a semiconductor than a conductor like mercury, the picture of the interphase region developed through the study of the mercury electrode is certainly a good place to start for an understanding of the cellular interphase. In Chapter 24, a somewhat different model of the double layer will be described for membranes whose only charge carriers are electrolytes. This model is called the *interface of two immiscible electrolyte solutions (ITIES).*

What happens when two phases are brought into contact with one another? Perhaps an easy way to consider this question is to take a short thought trip. Consider, first, the environment experienced by a single water molecule moving from a phase of pure bulk water to the interface with a phase of benzene (Fig. 20.1). Initially, deep inside the aqueous phase, the water molecule is free to look around and find the monotonous uniform tug (at least on a time-average scale) of other water molecules. No matter which way it "looks" it will "see" (experience forces exerted on it) the same thing. Specifically, there will be no net dipole orientation,

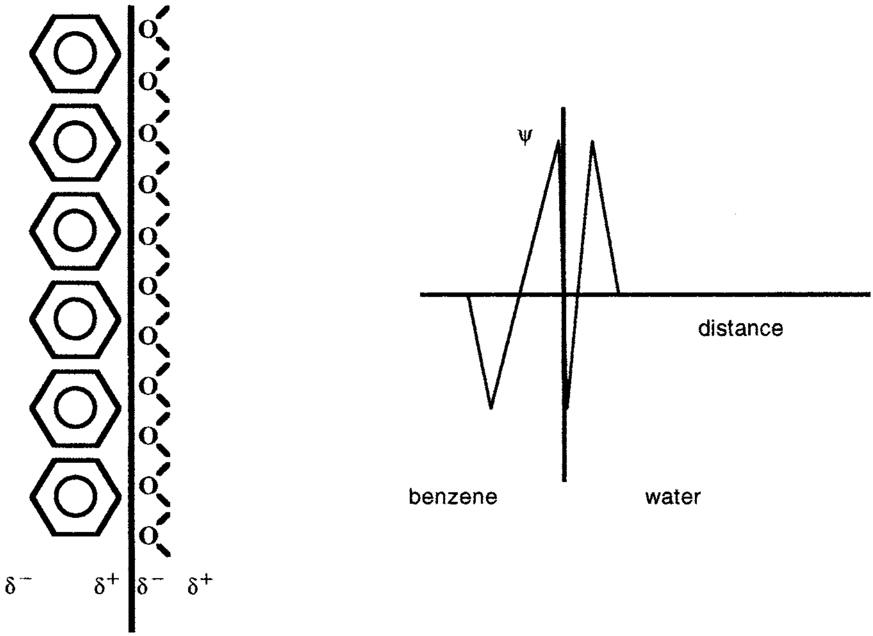


Fig. 20.1 The electrification of the interphase region due to the orientation of charges between the two phases, water and benzene. The *graph* is a representation of the change in potential, ψ, with respect to the distance into each phase

and electroneutrality will prevail in any reasonably chosen sample of the bulk phase. For the convenience of the argument, a lamina is chosen parallel to the benzene–water interface. Now, if the water molecule was to move randomly toward the interface between the benzene and water and suddenly was to find itself looking out not on another watery neighbor, but instead now on a benzene molecule, it would be profoundly changed in its view of the world. It has gone from an *isotropic environment* where direction is inconsequential to an *anisotropic environment* where the forces experienced depend on magnitude as well as direction. In fact, the water molecule will begin to experience a gradient of altered forces produced by the benzene–water interface before coming face to face with a benzene molecule as neighbor. The behavior of the molecules that experience the anisotropic environment will no longer appear random on a time-average scale. Water molecules located near the interface will experience orienting forces that depend on the nature of the interactions with the benzene molecules. In this case, it would be found that the water molecules will arrange themselves with their oxygen atoms facing the boundary, and their hydrogens facing into the aqueous phase. The layer of water apposite to the benzene phase, therefore, is no longer randomly oriented, and the arrangement of the water dipoles results in a lamina of charge. Because the water dipoles are no longer randomly oriented, as is the condition in the bulk phase, a charge separation has occurred in association with the interface. The charge derived from the aqueous dipoles is sensed across the phase boundary, and the benzene phase responds by producing a countercharge of equal but opposite magnitude. (In fact, the thought experiment could have been considered from the point of view of the benzene molecules with focus on a different mechanism for generating the benzene charge but yielding the same overall result.) Across the phase boundary, a potential difference exists, usually with a magnitude of no more than 1 V. However, the distance of the separation is generally quite small (1 nm) and therefore the resulting electric field has a magnitude of approximately 10^8–10^9 V/m. Such pairs of separated charges across the boundary between two phases constitute what is called the *electrical double layer* or the *electrified interface*. The region affected by this electric field extends not only across the phase boundary but also for a distance into each phase. In this region, the forces acting on the molecules are significantly different from those in the bulk of either phase, and the behavior of the molecules therefore deviates significantly from that of their bulk phase. The region where such deviation from the bulk properties occur is defined as the *interphase*.

The interphase includes both sides of the phase boundary, and if the entire interphase is considered, the electroneutrality principle will prevail. As previously mentioned, the existence of this electrified interface is a completely general phenomenon when phases meet. Potentials are developed by dissimilar metals in contact at liquid–liquid junctions, at liquid–solid junctions, etc. Even when a metal is in contact with a vacuum, there is a tendency for the electrons to protrude out of the metal surface, leading to a negative charge above the boundary countered by a positive layer inside the surface.

A range of mechanisms lead to charge separation at phase boundaries, and while not all have equivalent relevance in biological systems, some of the mechanisms will be presented here by way of a preview before a more detailed consideration of the interphase is undertaken. These mechanisms can be separated into categories that are somewhat artificial but can nonetheless be valuable in understanding the forces that lead to electrification of interphases (Table 20.1).

Table 20.1 Classification of the variety of mechanisms leading to interphase electrification

Type I charge established or imposed on one phase
Ionizable surface groups
Charged metal electrode surfaces
Physical entrapment of charge
Type II charge generated by differential affinity between phases
Difference in electron affinity between phases
Preferential attraction of ionic species to surfaces
Differential distribution of ions between phases
Arrangement of permanent dipoles to generate net charge separation
Arrangement of inducible dipoles to generate net charge separation

The interactions categorized as type I are similar in that one of the phases contains an identifiable charge when compared to the other phase. This should not be taken to suggest that the so-called charged phase is not electroneutral but rather that definable charged entities (ions and electrons) can be identified as associated with a specific phase whether or not the interface exists. For example, a surface containing carboxyl groups will become charged when the pH of the environment surrounding these groups rises above their pK_a. Although the development of this surface ionizable charge clearly depends on the pH of the associated phase, the charge must remain resident on the surface of the phase that contains the carboxyl groups. This particular case is especially applicable to biological systems and includes the ionizable charges associated with cell membrane components, the extracellular matrix, and the cellular glycocalyx. The entrapment of charge in a phase can occur in cases where the charge is associated with a component that cannot diffuse through a semipermeable membrane. This can also be an important effect biologically since many membranes are impermeable or variably permeable to different cations and anions. When a metal electrode is connected to an external power source, electrons can be either forced into or removed from the electrode, resulting, respectively, in a negatively or a positively charged electrode. The charge associated with the altered electron density of the electrode secondary to the external source is different from a charge separation that also results from the interaction of the electrode phase with its contiguous phase. The treatment of these two separate charges applies for all the type I. These forces will be examined in some detail shortly.

Type II mechanisms are invoked in cases where the charge separation results from an interaction that depends to a great extent on the interface being present. For example, the difference in affinity for electrons between two dissimilar metals leads to the contact potential difference between them. The charge separation is the result of the greater attraction of conductivity band electrons from one metal into the other metal. This case is probably more important in metal–metal and metal–semiconductor junctions than in metal–solution systems. In cases where an aqueous electrolyte constitutes one of the phases, as in all biological systems, specific anionic and cationic species will be adsorbed on surfaces in a differential fashion leading to charge separation. This chemically specific adsorption is often referred to as *specific adsorption.* Anions and cations will also partition across a phase boundary in a differential manner because of preferences for one phase or the other. A special case of this is the difference in potential, called the *diffusion potential,* which results when two similar solutions (but of different concentration) are brought together, and due to differences in the rates of diffusion of the ions, a potential difference results. The diffusion potential will be discussed further in Chapter 24. The case of phases that contain permanent dipoles, or may be induced to contain dipoles, has already been presented.

20.2 A Detailed Structural Description of the Interphase Is a Task for Physical Study

If an electric field is to be expected at every surface and interface between two phases, what will be the arrangement of the molecules and forces interacting in this region? To answer this question, an examination of the situation at the metal–electrolyte interface is valuable since it is reasonably well described. The principles can be generalized to situations relevant to the biological scientist, marking a solid starting place for the discussion.

A highly polished metallic electrode with the geometry of a flat plate is brought into contact with an aqueous solution containing only a true electrolyte (consideration of organic materials in solution will come later). The bulk electrolyte solution has the behavior already described in earlier chapters. The electrode can be connected to an external power source that will add electrons to or withdraw electrons from the electrode so that the electrode may have a negative, a positive, or no charge associated with it. The metal phase will be considered as a crystalline lattice of positive ions that are immobile with a cloud of electrons free to move in response to an electric field. An excess charge on the electrode, whether negative or positive, can be considered to be localized to the surface of the metal or to the electrode surface of the solution–electrode interface. The setup is depicted in Fig. 20.2.

A positive charge is placed on the electrode. Except for the geometry, the electrode now appears to the ions and water molecules making up the bulk electrolyte as a large cation. As in the case of the Debye–Hückel central ion, the electrolyte

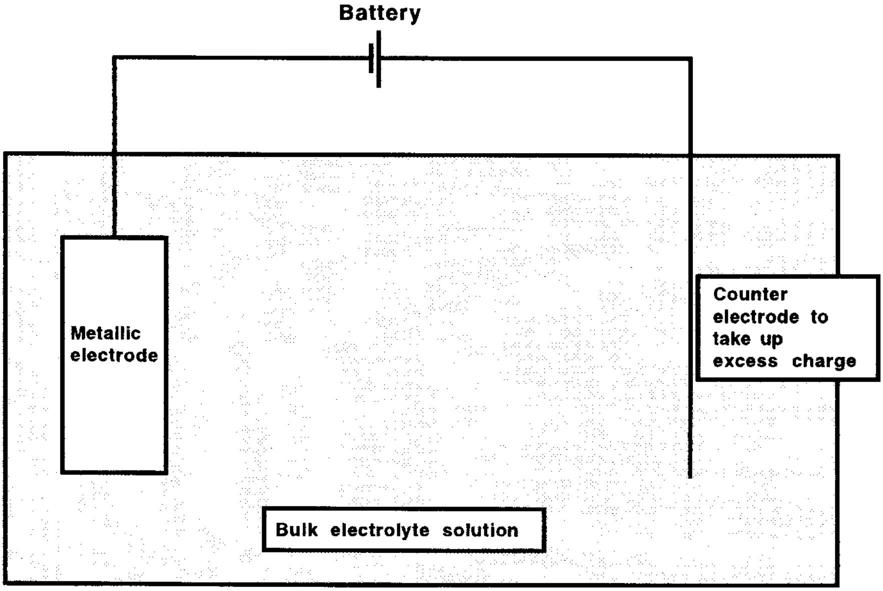

Battery
Metallic electrode
Counter electrode to take up excess charge
Bulk electrolyte solution

Fig. 20.2 System for the study of electrode–electrolyte interactions

will respond by orienting its molecules in such a way as to neutralize the electrode charge. In other words, a cloud of countercharge will form parallel to the electrode surface, similar to the one formed around the central ion. By moving molecules so as to counteract the charge on the electrode, the electrolyte phase itself now becomes charged in the region of the electrode. Sufficiently far into the bulk phase of the electrolyte, the charge on the electrode is no longer felt and the charge distribution in the solution phase becomes independent of the electrode charge. The properties of the interphase end and those of the bulk phase take over. What is the arrangement of the molecules making up the solution phase in the interphase region?

The structure of the ionic side of the interface has been, and continues to be, an area of intense study. There is good agreement at this time as to the structural elements of the charged ionic phase, due in large part to the work of Grahame. Rather than a highly detailed and mathematical treatment of the interphase region, a historical and qualitative approach will be used to describe the arrangement. Appreciation of the historical aspects of the development of the model of the aqueous portion of the electrified interface or double layer is useful because a great deal of the terminology relating to the science of the interphase is historically derived. It is also worth noting that similarities between the structure of electrolyte solutions around electrodes and ions were not lost on the researchers working on these problems in the early part of this century. In fact, as mentioned in Chapter 16, it was the work of Gouy in 1910 and Chapman in 1913, over a decade before Debye and Hückel's central ion model, that suggested the idea of the ionic cloud as a model.

20.3 The Simplest Picture of the Interphase Is the Helmholtz–Perrin Model

Just as the molecular structure around an ion was found by determining the charge density of the ionic cloud with respect to the distance *r* from the central ion, the structure of the electrified interface can be considered by evaluating the location and density of charge excesses with respect to the distance from the electrode. In a fashion similar to that for the ion, the parameter describing the charge excess is the potential resulting from the charge separation. The charge associated with a metal electrode has been described above as localized in a single layer on the surface of the electrode. The mission of the aqueous electrolyte is to neutralize this charge, and in the earliest model of the electrified interface, developed independently by Helmholtz and Perrin at the turn of the century, the entire countercharge to the electrode was considered to reside in a single rigid layer of counterions. This compact and immobile layer of ions was thought to completely neutralize the charge on the electrode. Were the model of the interphase to stop at this point, the electrified interface would be precisely analogous to a parallel plate capacitor composed of two plates of opposite charge, a double layer system. This is the historical derivation of the term *the double layer*. While the Helmholtz model is generally not adequate for describing

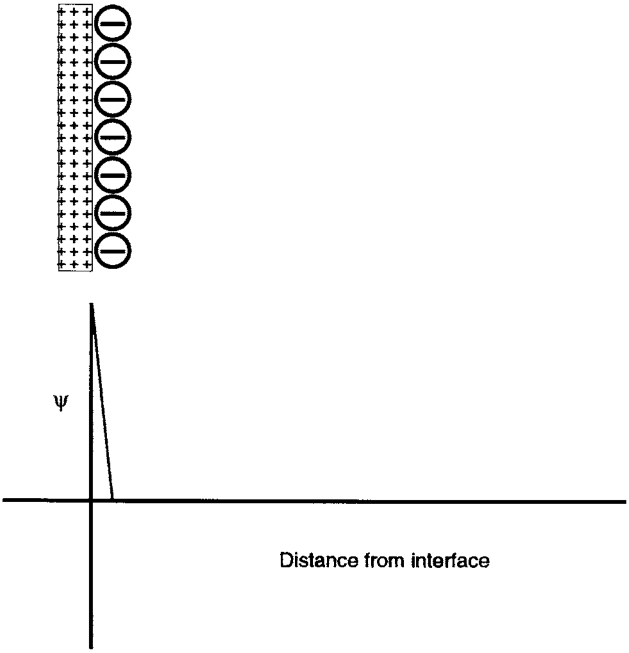


Fig. 20.3 Structure and associated potential versus distance curve of the Helmholtz model of the electrical double layer. Note that the arrangement of counterions forms a parallel layer that acts like a parallel plate capacitor, and hence the potential curve is linear

the electrified interphase region, the term has remained. If the Helmholtz model of the double layer were correct, then, since the charge on the electrode is completely neutralized by a single plane of ionic charge, a graphical plot of potential versus distance from the electrode would be linear as shown in Fig. 20.3.

It should be noted that no mention has been made of the arrangement of water and its structure in the interphase so far. Water plays an important structural role in the interphase and will be considered shortly, but the Helmholtz–Perrin model considers only electrostatic type interactions, and therefore only the ion–electrode relationships are examined here.

20.4 The Balance Between Thermal and Electrical Forces Is Seen as Competition Between Diffuse-Layer Versus Double-Layer Interphase Structures

One of the fundamental problems with the Helmholtz–Perrin model lies in its lack of accounting for the randomizing thermal effects ever present in liquids. Gouy and Chapman suggested that the capacitor plate-like arrangement of counterions be replaced by a diffuse cloud of charge that was more concentrated near the electrode surface and extended out into the bulk solution. Their formulation was based on Maxwell–Boltzmann statistics and is similar to the picture already developed for the Debye–Hückel cloud relationship of charge density in the cloud to the potential, ψ_r from the ion. In the *Gouy–Chapman model*, the decay of the potential derived from the electrode versus the distance, x, into the bulk electrolyte depends in part on the charge, z_i, on the ion and on the ionic strength of the solution. In fact, a thickness for the Gouy–Chapman diffuse layer exactly analogous to the reciprocal length parameter, κ^{-1}, of the Debye–Hückel model can be described:

$$\kappa^{-1} = \left(\frac{\varepsilon_{\rm o} \varepsilon kT}{2N_{\rm A} \left(z_i^2 \varepsilon_{\rm o} \right)^2 I} \right)^{1/2} \tag{20.1}$$

This means that the charge contributed by the diffuse layer can be simulated and treated as if it were a capacitor plate placed in parallel, a distance from the electrode equal to the reciprocal length. As in the Debye–Hückel model, only electrostatic interactions are considered; the ions are treated as point charges, and the solvent is considered as a structureless continuum. There is no consideration of the concept of closest approach of an ion to the electrode, ion size, or water interaction. The structure and potential versus distance relationships of a “pure” Gouy–Chapman diffuse layer are illustrated in Fig. 20.4. ψ varies with respect to distance, x, from the electrode according to the following equation:

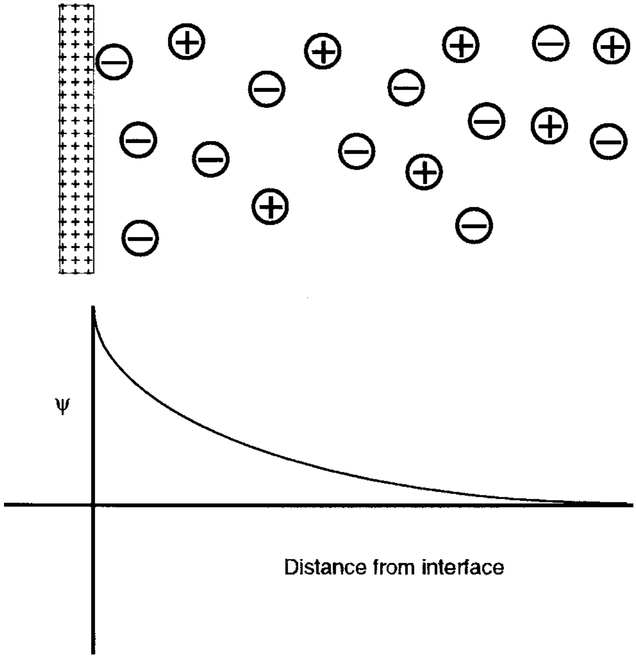


Fig. 20.4 Structure and associated potential versus distance curve of the Gouy–Chapman model of the interphase region

$$\psi_x = \psi_0 e^{-\kappa x} \tag{20.2}$$

where $\psi_x \rightarrow \psi_0$ as $x \rightarrow 0$.

20.5 The Stern Model Is a Combination of the Capacitor and Diffuse Layer Models

The Gouy–Chapman treatment of the double layer ignores the effect on the dielectric constant of the high potential fields (approximately 10^8 V/m) present at the interface, the fact that ions have finite size, and the tendency of molecules to adsorb on surfaces through forces other than electrostatic interactions (specific adsorption). In 1924, Stern presented a model that incorporated the questions of finite size and adsorption onto the surface into the structure of the interphase. The mathematical approach of Stern will not be discussed, but a qualitative picture can be sketched. When the finite size of an ion (whether hydrated or not) is considered, it becomes obvious that the distance between a countercharge comprised of ions and an electrode will be greater than if the countercharge were comprised of point charges. The inclusion in the model of interactions at the surface similar to those described by the Langmuir isotherm added an important dimension for the molecules that are on the inside of the interphase region. Because of the adsorptive forces (often described as the "chemical" aspect of the interaction) acting on certain ionic species, the number of ions, and hence the charge excess, in the layer closest to the electrode will be altered when compared to that predicted by a model in which only electrostatic interactions are taken into consideration. The number of ions populating the innermost layer, where adsorption is important, and the distance of closest approach of this layer to the electrode are both finite. Because there is a limit to the nearest approach of the ion cloud, a region of linear potential decay with respect to distance results. Such a region has the behavior of the molecular capacitor suggested by Helmholtz. This inner layer is called the *Stern layer*. Beyond the inner layer, the Stern model predicts that the remainder of the charge needed to neutralize the electrode charge is arranged like a Gouy–Chapman diffuse layer. The *Stern–Gouy–Chapman model* is illustrated in Fig. 20.5. A very important point that derives from the Stern model is that the amount of charge excess found in the Stern layer or the diffuse layer differs under various conditions of electrolyte valence and concentration. In dilute solutions, the countercharge will be found almost completely in the Gouy–Chapman diffuse layer, while in solutions sufficiently concentrated the charge may instead be localized almost completely on the Stern layer, giving rise to a Helmholtz-type structure.

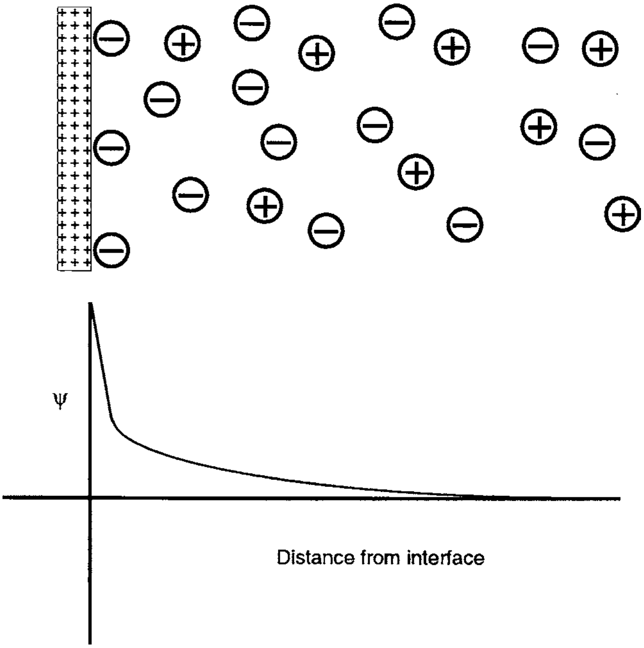


Fig. 20.5 Structure and associated potential versus distance curve of the Stern model. Note that the model contains elements of both the Helmholtz and Gouy–Chapman formulations

20.6 A More Complete Picture of the Double-Layer Forms with Added Detail

Further developments have been made with respect to understanding the structure of the double-layer region, and the role of water and its dielectric properties have been taken into account. Building on the historical basis laid to this point, a modern picture of the double-layer region will now be drawn (Fig. 20.6). To develop this picture, imagine moving from the surface of the metal electrode (actually the starting point will be just inside the charge layer of the electrode). As in the model for the ion developed earlier, water molecules will be the first components of the solution encountered. A monomolecular layer of water molecules will be lined up immediately adjacent to the electrode surface. In the case of the positive electrode, the negative end of the water dipole will be preferentially facing the electrode, but this orientation will reverse itself if the charge on the electrode is reversed. The electrode will be hydrated in a fashion analogous to that for an ion. Unlike an ion, however, the metal electrode can be forced to change its charge, and the direction of the water dipole can also be induced to change to a varying degree with the electrode charge. The arrangement of the dipoles and the energy of their attraction to the electrode have been shown to play an important role in allowing or preventing adsorption of organic molecules on electrodes.

On moving through the hydration sheath, a second layer of molecules is encountered, comprised primarily of hydrated ions of appropriate countercharge. In the

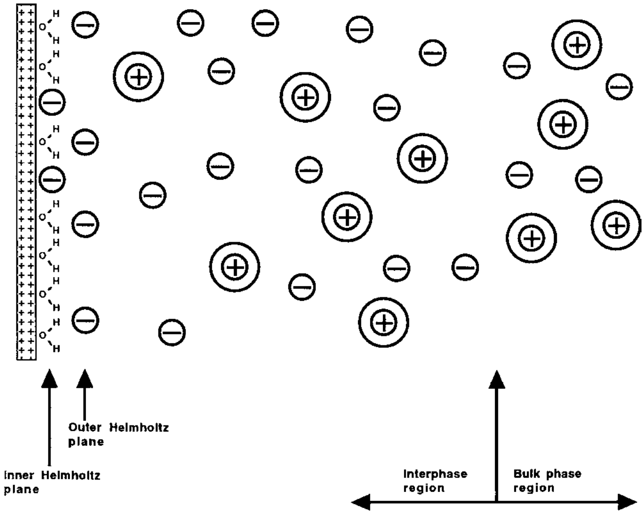


Fig. 20.6 Schematic of the structure of the interphase, including orientation of aqueous dipoles, specific adsorption of unhydrated anions on the electrode surface, and hydrated cations. Note that the outer Helmholtz plane is the division between the Helmholtz-like molecular capacitor and the diffuse Gouy–Chapman layer

case of the positive electrode, this layer is populated by anions, and for the negatively charged electrode, cations. This layer of ions, an important constituent of the molecular capacitor, is called the *outer Helmholtz plane* (OHP). The outer Helmholtz plane is the dividing line in the modern view between the outer diffuse Gouy–Chapman layer and the inner compact region of the interphase. Attention will be turned now to the structure of the inner compact region. The electrode is considered to be covered with a monomolecular layer of water with a net dipole orientation determined by the sign of the charge on the electrode. In some cases, it is possible that ions can displace these adsorbed water molecules. When the electrode is positively charged, the counter field ions will be comprised primarily of anions. Anions differ from cations in their surface activity because they are usually not hydrated. Just as is the case with a water–air interface where the anions are found preferentially at the surface, anions will move in toward the hydration sheath of the positively charged electrode and displace a water molecule. The plane of closest approach to these unhydrated anions is called the *inner Helmholtz plane (IHP)*. Whether an ion will leave the outer Helmholtz plane and move to a position on the inner Helmholtz plane depends on the free energy change associated with the jump from one plane to the other. If the ΔH and ΔS values for the ion–electrode, water–electrode, and ion–water interactions are all considered, ΔG for the jump can be found. As Table 20.2 shows, ΔG for this jump is positive for Na^+ and K^+, while it is negative for Cl^- and I^-. Consequently, for these selected ions of biological importance, the anions will adsorb at the IHP while the cations will remain in the OHP. The specific adsorption of anions on the electrode surface occurs even when the electrode is negatively charged. The tendency of anions to be surface active, and therefore present at the inner Helmholtz plane when simple electrostatic forces would not predict their presence, leads to the interesting case shown in Fig. 20.7.

Table 20.2 ΔG values for the movement of selected ions to the inner Helmholtz plane[a]

	Interaction				
			Ion–water		
Ion	Ion–electrode ΔG	Water–electrode ΔG	ΔH	ΔS	Total ΔG
Na^+	−207.5	113.0	165.3	46.9	56.5
K^+	−204.2	79.9	141.8	21.7	10.9
Cl^-	−208.8	76.6	74.1	−69.5	−37.2
I^-	−210.5	69.0	45.6	−136.4	−54.8

[a] ΔG and ΔH values are in kJ/mol; ΔS values are in J/mol/deg

These concepts of interphase structure are important in a large number of phenomena associated with biological processes. In the next chapters, the effects of this structure on physiochemical behavior of macromolecular structure and membranes will be explored.

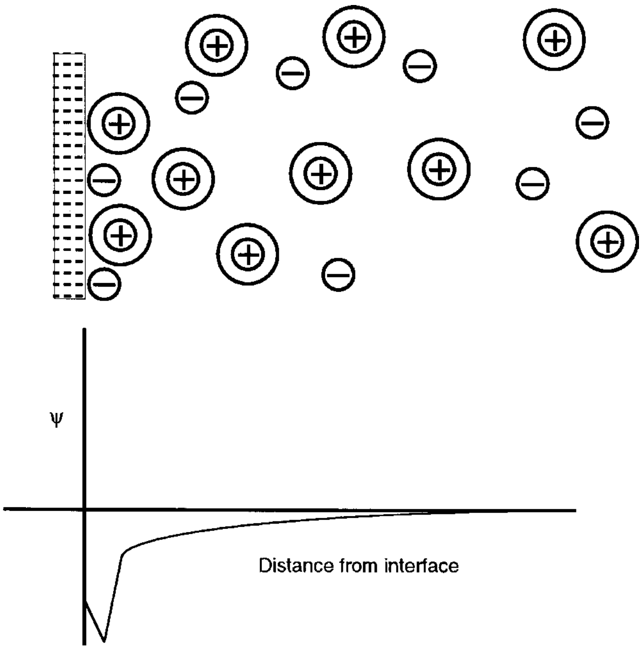


Fig. 20.7 Structure of the interphase and the potential versus distance curve for a case of specific adsorption of anions on a negatively charged electrode

20.7 Colloidal Systems and the Electrified Interface Give Rise to the Lyophilic Series

Colloidal dispersions are comprised of particles whose size is on the order of 10^{-6} to 10^{-9} m, and consequently an enormous surface area per unit volume is exposed. The large surface area means that surface behavior is a powerful and frequently dominant property of these systems. The behavior of colloidal systems must be interpreted to a large extent on the properties and phenomena deriving from the electrified interface. Macromolecules and organelles in the cell have many of the characteristics of colloidal systems. Therefore, the behavior of the cells is strongly influenced by the interphase properties of the electrified interface.

Colloids are frequently categorized as either lyophilic or lyophobic. Lyophilic refers to colloids that are considered "solvent-loving" and lyophobic to those that are "solvent-hating." *Lyophobic colloids* form systems that are called sols, and *lyophilic colloids* form systems called gels. Many aspects of cellular and biochemical systems involve the behavior of sols and gels. Gels have in fact already been extensively covered in this text, since solutions of macromolecules are three-dimensional gels. Many lyophilic colloidal materials can be formed into two-dimensional networks or

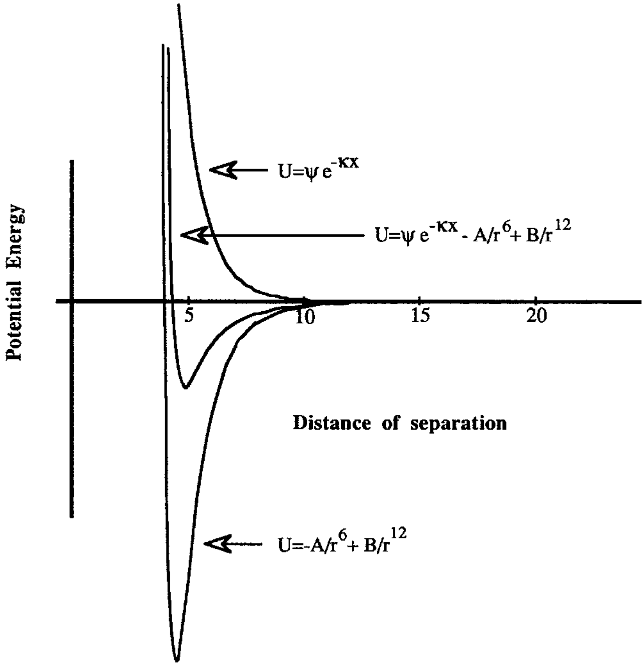


Fig. 20.8 Plot of the interactional force between two colloidal systems, each having a double layer of identical charge. The *top* curve represents the energy of interaction of the double layers. The *bottom* curve is a Lennard–Jones potential curve for the system, and the *middle* curve is the interactional energy that results from the superposition of the *upper* and *lower* curves

matrices. These gels are comprised of intertwined lyophilic molecules in a continuous mass with fine pores dispersed throughout the gel, constituting a membrane. The behavior associated with the pores in these membranes is greatly dependent on double-layer effects (as is the case with electro-osmosis in ion-exchange membranes). The cytosolic dispersion of organelles and suspensions of many cells including bacteria have properties similar to those of sols. A knowledge of colloidal systems will often help in understanding cellular and biochemical behavior.

Lyophobic colloids, such as colloidal gold or droplets of lipid in an aqueous solvent, exist as a dispersion of particles suspended in the solvent. The formation of micelles as described in Chapter 22 forms dispersion of colloidal ions. Interactions of cells may be reasonably considered using these colloidal particles as models. It may be correct to treat organelles in the cytosol with the same model, although the properties of the cytosolic "solvent" are much less clear in this case. The particles prefer not to associate with the solvent and, if conditions are right, will associate with one another, and flocculation or coagulation occurs. Whether a lyophobic colloid exists in a dispersed or flocculated state depends greatly on the presence of the double layer. What are the forces acting on a pair of identical lyophobic particles as they approach one another in the dispersed state? Since each of the particles is surrounded by an identical electrified interphase region, it is this charge layer that each particle sees as it approaches the other. The force that exists between the two particles is an electrostatic one and, since the charges are the same, the force will cause the particles to repel one another. The potential energy of the interaction will increase progressively as the double layers of the particles approach one another. Because it is the two double layers that are interacting, the electrostatic force is approximated by

$$F_{\text{double layer}} = U_{\text{electrostatic}} = \psi_0 e^{-\kappa x} \tag{20.3}$$

A graph of this relation is shown in Fig. 20.8.

As the particles approach one another, other interactive forces come into play, given by a curve like the Lennard–Jones potential (cf. Chapter 9). Earlier, this curve was described as including interactions due to the attractive van der Waals forces and repulsion of the electron cloud:

$$U_i = -\frac{A}{r^6} + \frac{B}{r^{12}} \tag{20.4}$$

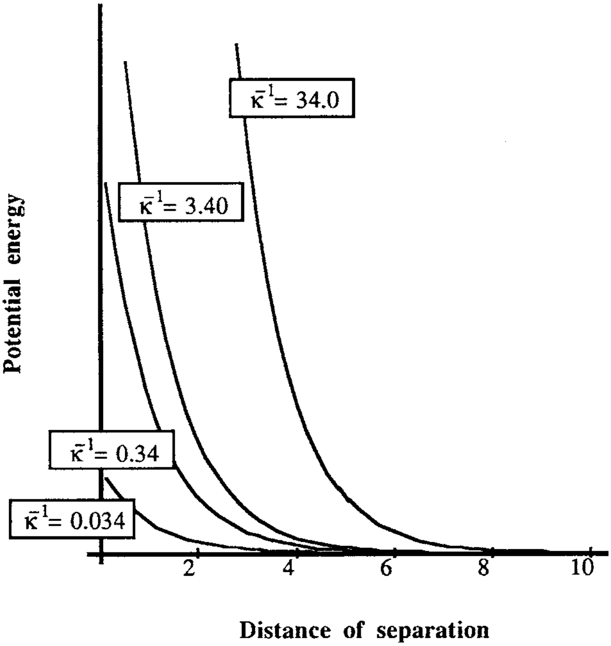


Fig. 20.9 Plot of a family of curves relating the double-layer thickness to the interactional energy with respect to distance

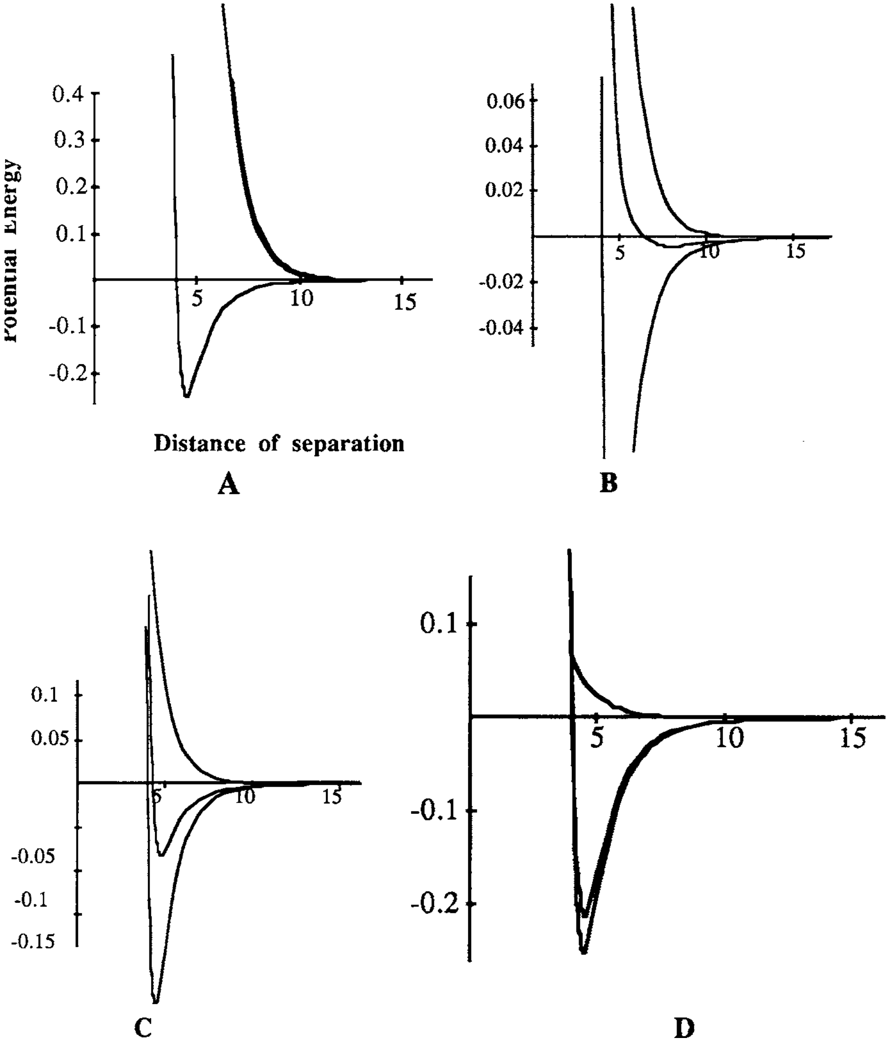


Fig. 20.10 Potential energy plots for the interaction of colloidal systems, showing alterations that are caused by the adsorption of high-valency ions at the inner Helmholtz plane. In each panel, the *bottom* curve is a Lennard–Jones curve, the *top* curve is the interactional energy between the double layers, and the *middle* curve is the superposition of the upper and lower energies. In panel A, ψ_0 is given an arbitrary magnitude of 100. The interactional energy corresponds to a case where there is no IHP neutralization of charge, and the electrostatic field is quite strong for a significant distance into the interphase region. The colloidal systems are therefore unable to have an attractive interaction. In panel B, the adsorption of counterions at the IHP has decreased ψ_0 to a magnitude of 20, and a quite small potential energy minimum is found. Panels C and D represent the reduction in ψ_0 to 5 and 1, respectively, and show dramatically the effect that increasing the valency of the counterion can have on the interactional energy of the system. As the electrostatic field is reduced in intensity, the attractive interactions at close distance become much more favorable and association can easily occur

The Lennard–Jones potential curve is redrawn in Fig. 20.8. The complete potential energy curve for the interaction of the lyophobic particles is given by combining the electrostatic interactions with the Lennard–Jones potential:

$$U_{\text{colloid}} = \psi_0 e^{-\kappa x} - \frac{A}{r^6} + \frac{B}{r^{12}} \tag{20.5}$$

As can be seen in Fig. 20.9, when the thickness of the double layer is sufficient, the potential energy curve of interaction will not favor the approach of the particles. They will stay dispersed under these conditions.

However, the thickness of the double layer can be made to vary by altering the ionic strength of the electrolyte, and the electrostatic potential curve can be moved. As can be seen from the family of curves in Fig. 20.10, increasing the ionic strength, reflected by changes in κ^{-1}, moves the double layer in toward the surface of the particle and leads to a significant change in the interactional potential energy with respect to separation distance. As the double layer becomes more compressed, the van der Waals attractive forces become more important, and the potential energy for the flocculation of the dispersed particles becomes favorable. Hence, by changing the thickness of the double layer, the behavior of the colloidal dispersion is affected. Changing the ionic strength can be accomplished both by changing the concentration of a particular type of electrolyte, for example, increasing or decreasing the concentration of a 1:1 salt, or by changing the types of electrolytes that go into making a solution, for example, changing from a 0.10 M solution of a 1:1 electrolyte to a 0.10 M solution of a 1:2 or 2:2 electrolyte.

Manipulation of the double-layer thickness by varying the concentration can be cumbersome, and that is probably not the method of choice in a biological system. Another type of manipulation of the electrified interface can lead to significant changes in the interactive potential energy curve. Specific adsorption of ions at the inner Helmholtz plane can lead to a reduction in the magnitude of ψ_0, and hence the potential energy of interaction can be made to favor flocculation through the addition of these ions to the system.

20.8 Salting Out Can Be Understood in Terms of Electrified Interphase Behavior

It is now instructive to reflect on the precipitation of a colloid from solution. The precipitation of a lyophilic colloid was discussed in Chapter 15. In the discussion of solutions of ions, it was shown that when the activity of the aqueous solvent was reduced by the addition of electrolyte to a solution of macromolecules (a gel), an oversaturated solution was produced with the subsequent precipitation of the macromolecules. This process is mechanistically explained by the structural consequences of removing the hydration layers from the colloidal material. Relatively large amounts of electrolyte must be added to the gel to salt out the protein. In this chapter, it has been shown that the addition of electrolyte can also lead to the precipitation or salting out of a lyophobic colloid, a sol. However, now the salting out occurs by changing the interactional energies that result from the double-layer structure. Relatively small amounts of electrolyte are required for these effects. The point to be remembered is that biological systems are comprised of both sols and gels, often existing side by side. The differences in behavior stem to a great degree from the differences in interaction with the aqueous solvent. In the case of the sol, water avoids interaction with the colloid for the reasons described in earlier chapters, and so the only forces acting to keep the sol dispersed are the electrical forces derived from the interactions of the double layers. For the lyophilic colloids, the water associates to a much higher degree with the solute, and stabilization of the dispersion occurs because of forces above and beyond those associated with the electrified interface. It is clear that knowledge of the relationships described here, complex as they are, is vital to a complete understanding of the cell.

Further Reading

General

Bockris J.O'M. and Reddy A.K.N. (1970) *Modern Electrochemistry*, Volume 2. Plenum, New York.

Bockris J.O'M. (1983) Teaching the double layer, *J. Chem. Educ.*, **60**:265–268. (A master teacher's account of his approach to the subject of the double layer and its structure.)

Koryta J. and Dvorak J. (1993) *Principles of Electrochemistry*, 2nd edition. Wiley, New York.

Colloidal Properties

Hunter R.J. (1992) *Foundations of Colloid Science*, Volumes 1 and 2. Oxford University Press, Oxford.

Hunter R.J. (1994) *Introduction to Modern Colloid Science.* Oxford University Press, Oxford.

Leiken S., Parsegian V.A., and Rau D.C. (1993) Hydration forces, *Annu. Rev. Phys. Chem.*, **44**: 369–395. (Presents the view that many important colloidal behaviors depend to a great degree on the water organization near the surface.)

Problem Sets

1. Assume that an isolated preparation of cytosolic organelles has been obtained and is sitting in a test tube on your bench. Describe and discuss two methods for precipitating these organelles based on their colloidal nature. Discuss when one method might be preferred over the other.

2. Describe a molecular model for the organization of water molecules around a cell. Include in your analysis the orientation, structure, dimensions, and composition of the region.
3. Is the Stern or Grahame model of the interphase to be preferred over the Gouy–Chapman or Helmholtz model in biological systems?
4. In experiments to investigate the protein structure of the nucleosome octamer, the isolated nucleosome is placed in a very high ionic strength solution. This step causes the dissociation of the DNA from the histones. Explain this in terms of the interactional energies of nucleosome formation.

Part IV
Function and Action Biological State Space

Chapter 21
Transport – A Non-equilibrium Process

Contents

21.1 Transport Is an Irreversible Process and Does Not Occur at Equilibrium

So far, the focus of this book has been on systems at equilibrium. At equilibrium, systems experience no net flux of heat, work, or matter. Classical thermodynamics treats these systems easily. As pointed out earlier, the greatest value of thermodynamics is that the behavior of a system can be predicted, even when the mechanistic details are not known. Homogeneous systems, at constant temperature and pressure, such as the solutions of electrolytes and macromolecules described so far, are comprised of molecules that individually experience a variety of forces, both orienting and randomizing. On an instantaneous timescale, this might lead to net movements of mass or energy; however, the time average of the forces leads to the steady-state condition of equilibrium. The activity of a component is the reflection of the time-average molecular forces acting in a system at equilibrium. There are cases in which the time average of a force or forces acting on a system results in the flow of material. When these events occur, transport phenomena results. Transport phenomena and the principles associated with non-equilibrium behavior are extremely important in biological systems because, as has already been suggested, true equilibrium states are achieved only in death. Steady-state systems, which have constant fluxes, are common. These systems are treated by non-equilibrium methods.

There are four phenomena associated with transport. These include diffusion, electrical conduction, heat flow or conduction, and fluid flow or convection. Each of these represents net movement in the direction of a gradient from a higher to

DOI 10.1007/978-1-4419-6324-6_21, © Springer Science+Business Media, LLC 2010

a lower potential. The gradients are due to differences in chemical potential, electrical potential, temperature, or pressure, respectively. All of these phenomena are important in biological systems though our primary focus will be on diffusion and electrical conduction. The general equation that applies to all forms of transport events is written:

$$J_x = -B\frac{\partial A}{\partial x} = -BF_A \tag{21.1}$$

This equation states that the flow of material in the x direction, J_x, is proportional by some constant B to the gradient of force of type A in the x direction. Similar equations could be written for each coordinate, x, y, or z.

Transport phenomena do not fall in the realm of the classical thermodynamics. It is possible to analyze and study these non-equilibrium processes mechanistically, that is, to calculate the forces on each molecule and then relate the combined actions of each and every molecule to the properties of the system. This approach can be used to provide a qualitative picture of the events that lead to transport. There are substantial problems with a strict mechanistic approach. The first problem lies in defining the actual forces that may be acting on a particular molecule. As will be seen, the dimensions of molecules are important parameters in calculating transport properties mechanistically. In a system of macromolecules, this can be a significant difficulty. In many cases, biochemists and biophysicists do not even know all the components that go into making a system, much less their dimensions. Furthermore, the approximations for studying the forces on moving objects such as Stokes' law, which is a mainstay of this approach, assume that the transport occurs in a medium that is a continuum. Such an assumption, especially in the case of aqueous media, ignores the forces that act between a component and its solvent and other components, leading to approximations that can be drastically at variance with reality. Ideally, a set of laws and equations parallel (or complementary) to those applied in the equilibrium studies already described can be found for cases where equilibrium is approached but not yet reached. Such a macroscopic set of empirically (or phenomenologically) derived descriptions of properties (such as transport or kinetic rates) could complement the molecular-mechanistic approach. The study of non-equilibrium or irreversible thermodynamics provides this effective phenomenological approach. The subject of irreversible thermodynamics is a detailed and relatively complicated subject and we will only touch on the general ideas and vocabulary here. Details can be found in the references listed at the end of the chapter.

21.2 The Principles of Non-equilibrium Thermodynamics Can Be Related to the More Familiar Equilibrium Treatment with the Idea of Local Equilibrium

We have stated that the principles of thermodynamics are universally valid. It was never stipulated that only systems at equilibrium could be treated. The reason that only equilibrium systems have been treated to this point has been one of definition and convenience. This occurred because some of the fundamental variables of state, namely, temperature, pressure, and entropy, were defined at equilibrium. They are more difficult to define during an irreversible or non-equilibrium process. Other variables of state do not suffer from this limitation and can be successfully used under any circumstances; these include volume, mass, energy, and amount of a component. Recognizing that variables like temperature and pressure are intensive, while volume, mass, and energy are extensive can help explain this difference. An intensive property was defined as one in whose evaluation a small sample was representative of the entire system. This only has meaning if a system is at equilibrium. Consider the example of two heat reservoirs of different temperature connected by a metal bar through which heat travels by thermal conduction. The flow of heat will be irreversible from the reservoir of greater temperature to the one of lower temperature. Choosing small samples at points along the bar will give different measurements for the temperature. Clearly, this does not fit the requirement for an intensive variable. Consequently, the method used to define the parameter of temperature must be different in this system, because it is not at equilibrium. A similar argument could be made for pressure or entropy. Until the variables, such as temperature and pressure, can be defined in an irreversible system, thermodynamic calculations will not be successful.

Through the use of a new postulate, that of *local equilibrium*, this problem can be overcome. The system is divided into small cells, small enough, that effectively each point in the system is treated, but large enough so that each cell contains thousands of molecules. At a specific time, t, the cells are isolated from the system and allowed to come to equilibrium in time dt. Therefore, at time $t + dt$, measurements can be made that give an equilibrium temperature or pressure. The variable of state at time t is then considered to be equal to the measurable variable at time $t + dt$. The relationships derived by this postulate are then considered to be equivalent to the relationships derived from equilibrium states. It must be realized that this postulate has its limitations. The presumption is that the variables in the system are not changing too rapidly. If the time, dt, necessary for the cell to achieve local equilibrium approximates the time during which a change for the whole system may be measured, then the postulate cannot be reasonably applied

Entropy has been shown to play an important role in our systems of interest and will be seen as a crucial driving force in transport phenomena. How does the treatment of irreversible systems work in the case of entropy? Instead of employing the relationship $\Delta S = q_{\text{rev}}/T$, it is more convenient to determine ΔS from another relationship, for example

$$\Delta S = C_p \ln \frac{T_2}{T_1} \tag{21.2}$$

Once the entropy of each cell is determined, the free energy for each cell can be determined:

$$dG = VdP - SdT + \sum \mu_i dn_i \tag{21.3}$$

We know that reversible processes take an infinite amount of time to complete but do not lead to the production of entropy. Irreversible processes, on the other hand, occur in a finite time and create entropy. The rate of a process therefore can be defined in terms of the rate of entropy production with respect to time, $\frac{dS}{dt}$. This means that as a reaction or process proceeds in an isothermal system, there will be heat flow into and out of the surroundings and system. The differential change in entropy will be given by

$$dS = d_iS + d_sS \tag{21.4}$$

where d_iS is the entropy change in the system, and $d_{\underline{\mathrm{s}}}S$ that in the surroundings. dS will always be zero or greater.

Historically, the formulation of irreversible thermodynamics started when Thomson (Lord Kelvin) was investigating the relationship between the flow of electricity and heat flow in thermocouples. If two dissimilar metal wires are twisted together at one end and a voltmeter is used to complete the circuit between the two, a voltage can be demonstrated arising from the contact of the two phases. This is the *contact potential*. If both ends are connected together, there will be two junctions in the circuit. If these two junctions are isothermal and an electric current is passed between them, heat will be absorbed from the surroundings at one junction and an equal amount of heat will be released at the other junction. This heat flow is reversible in that when the direction of the current is changed, the direction of heat flow also changes. This reversible heat is called the *Peltier heat*. A second source of heat is also produced during this process due to the resistance of the metal to the flow of charge, and this heat is called the *Joule heat*. Joule heat is irreversible. If the two junctions are now placed at two different temperatures, an electromotive force will exist between them. The electromotive force between the two junctions is called the *Seebeck emf*. If a charge is allowed to move around the circuit because of the Seebeck emf, it will be found by experiment that the Peltier heat appearing at the junctions is not sufficient to account for the work accomplished. Thomson therefore proposed a second reversible heat associated with the flow of current, the *Thomson heat*. The Thomson and the Peltier heats are reversible and are accompanied by two irreversible heats in this system, one due to Joule heating and one due to heat conduction. Thomson treated the thermocouple as if it were a reversible heat engine in which only the Thomson and Peltier heats circulated. He described a series of relationships that showed that there was no entropy production, that is, that the two heats were equal and reversible. Thomson's theoretical treatment of this system was experimentally validated even though he ignored the two irreversible terms. Thomson himself recognized that the treatment was incomplete since the process described is an irreversible one and hence the total entropy of the process must be positive. However, his analysis assumed that the entropy increase associated with the Joule heating and the heat conduction would be positive and constant and therefore tested the hypothesis that the Peltier and Thomson heats did not add to the entropy generation of the process, that is, that they were indeed reversible. His result demonstrated that in transport phenomena there are reversible and irreversible processes

A unifying method for generally treating irreversible systems was given by Onsager in 1931. Onsager based his formulation on the *principle of microscopic reversibility*, which says that at equilibrium any process and its reverse process are taking place on the average at the same rate. He further assumed that for a process that is near equilibrium, equations may be written for the transport process in which the fluxes are linearly proportional to the forces. The theory is valid only for deviations from equilibrium where this linear relationship exists. Processes like the one described above can be generally treated by considering that in a transport process there are a number of flows that occur simultaneously. For example, in the case of thermoelectricity, there is a flux of heat, $J_{1,}$ and one of current, J_2. The two flux equations take the general form:

$$\begin{aligned} J_1 &= L_{11}X_1 + L_{12}X_2 \\ J_2 &= L_{21}X_2 + L_{22}X_2 \end{aligned} \tag{21.5}$$

The term X_x represents the force gradient, L_{ij} are the phenomenological coefficients, and L_{ii} are the direct coefficients. In this case, X_1 represents the temperature gradient and X_2 the electrical gradient. The forces represented by X_x are thermodynamic driving forces and have the form

$$\frac{\partial S}{dX_X} = F_x \tag{21.6}$$

This type of analysis indicates that when more than one gradient is causing flux, there will be a coupling of the flows. The direct coefficients represent the fluxes due to the directly related gradient, that is, the flow of heat due to a thermal gradient. These always increase the entropy of the reservoirs. The cross terms, L_{ij}, are the coupled flows caused by the gradient that is not directly related, that is, the flux of heat caused by the flow of electricity. Onsager showed that the phenomenological coefficients are equal:

$$L_{ij} = L_{ji} \tag{21.7}$$

This equality is called the *Onsager reciprocity relation*. The coupling between the flows indicates the interaction of one flow with another. The idea that some fluxes are independent while others occurring simultaneously are interacting and reversibly coupled is an important one in transport phenomena.

In mechanistic terms, transport can be thought of as the balance between the motion of a particle moving directly down a gradient of force and the scattering of the particle away from this direct path because of interactions between other forces or particles. While the relationships of irreversible thermodynamics are probably the most accurate expression of the balance between direct and scattering forces, there are two mechanistic models frequently considered. One is the concept of the mean free path, and the other is based on relaxation time. In the next chapter we will focus on the concept of the mean free path and see how this approach can be used to examine diffusion in a solution

Further Reading

Haase R. (1969) *Thermodynamics of Irreversible Processes*. Dover Publications, New York.
Waldram J.R. (1985) *The Theory of Thermodynamics.* Cambridge University Press, Cambridge.

Chapter 22
Flow in a Chemical Potential Field: Diffusion

Contents

22.1 Transport in Chemical, Electrical, Pressure, and Thermal Gradients Are All Treated with the Same Mathematics

Transport is measured by *flux*, which is defined as the net movement of matter in unit time through a plane of unit area normal to the gradient of potential. In the case of *diffusion*, molecules of specific components will travel down their concentration gradient, $\partial c/\partial x$, until the gradient is removed or neutralized. During *electrical conduction*, mass, as well as charge, is transported in an electric field $\partial \psi/\partial x$. In aqueous solutions, the mass and charges are derived either from electrolyte ions or from the autoproteolysis of water which yields protons and hydroxyl groups. In *convection* or fluid flow, molecules are moved by a streaming process caused by an external force $\partial P/\partial x$, for example, pressure from a pump as in the cardiovascular system or gravitational forces generated in a centrifuge. *Heat conduction*, caused by a difference in temperature $\partial T/\partial x$, is also the result of a net flow, not of the total number of particles, which remains constant in the system, but of the particles with a higher kinetic energy. Each of these transport phenomena has a named relation of the form given earlier by

$$J_x = -B\frac{\partial A}{\partial x} = -BF_A \tag{22.1}$$

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_22, © Springer Science+Business Media, LLC 2010

These named relations are

(1) for diffusion, it is *Fick's law* and the constant is given as *D*, the *diffusion constant*;
(2) for electrical conduction, it is *Ohm's law* with the constant κ, the *electrical conductivity*;
(3) fluid movement is given by *Poiseuille's law* and the constant, *C*, is a *hydraulic conductivity* related to the viscosity;
(4) heat flow is described by *Fourier's law* and κ_{T} is the *thermal conductivity coefficient.*

22.2 Diffusion or the Flow of Particles Down a Concentration Gradient Can Be Described Phenomenologically

The treatment of diffusion and chemical potential gradients can be illustrated by considering a rectangular container of an aqueous solution of *i*, as in Fig. 22.1. For simplicity, assume the solution is ideal and so the activity of *i* can be given in terms of the concentration of *i*. In this rectangular container, the concentration of *i* is allowed to vary in the *x*-direction only; that is, there are a series of planes, in the *yz*-plane, perpendicular to the *x*-axis such that the concentration of *i* is equal within each plane, but may vary from plane to plane. There will be a gradient in concentration measured moving along the *x*-axis from plane to plane. The concentration gradient can be described in terms of the chemical potential as follows. The first plane chosen for description will be at the position $x = 0$. The chemical potential at $x = 0$ can be described by the term

$$\mu_{i-0} = \mu_i^o + RT \,\mathrm{In}\, c_0 \tag{22.2}$$

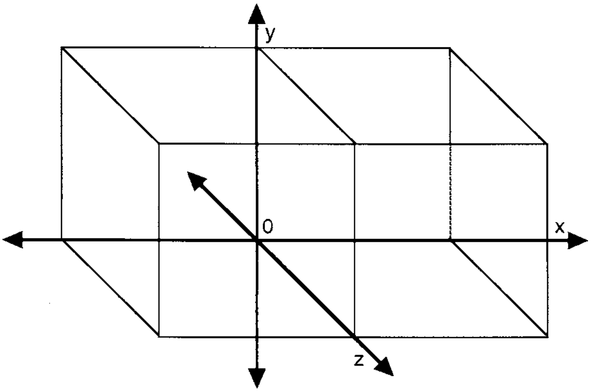


Fig. 22.1 Container system for the consideration of chemical potential gradients

At the end of the container, $x = f$, there will be a plane where the chemical potential can be given as

$$\mu_{i-f} = \mu_i^o + RT\,\mathrm{In}\,c_f \tag{22.3}$$

The change in free energy associated with moving a mole of species i from $x = 0$ to $x = f$ will be the difference in the chemical potential of the planes; thus

$$\Delta\mu_i = \mu_{i-f} - \mu_{i-0} = RT\ln\frac{c_f}{c_o} \tag{22.4}$$

From the earlier discussion of thermodynamics, it is known that the free energy change is equal to the net work done on the system in a reversible process at constant temperature and pressure. Therefore

$$\Delta\mu_\mathrm{i} = w \tag{22.5}$$

Work is defined as force acting over a distance and hence it could be written that the work and the free energy change are related to a force (of diffusion) acting over the distance between 0 and f:

$$w = -F_\mathrm{D}(x_f - x_0) = -F_\mathrm{D}\Delta x = \Delta\mu_i \tag{22.6}$$

The justification for the negative sign will be discussed shortly. The force, F_D, can be written in terms of the free energy change and the distance:

$$F_\mathrm{D} = -\frac{\Delta\mu_i}{\Delta x} \tag{22.7}$$

The gradient can be written in terms of infinitesimal changes, giving:

$$F_\mathrm{D} = -\frac{d\mu_i}{dx} \tag{22.8}$$

Equation (22.8) states that the force leading to diffusion is equal to the negative gradient of the chemical potential. The negative sign can be explained by analogy to the work done in lifting a weight in a gravitational field. If a weight is lifted from $y = I$ to $y = F$, then the difference in potential energy of the weight, U, will be given by

$$w = \Delta U \tag{22.9}$$

The work is done in a gravitational field so an expression analogous to Eq. (22.6) can be written as

$$w = -F_\mathrm{g}\,(y_\mathrm{F} - y_\mathrm{I}) = -F_\mathrm{g}\Delta y = \Delta U \tag{22.10}$$

The gravitational force is acting in a downward direction and, because of its vector, will be negative; the displacement of the weight will be upward and so Δy will be a positive quantity. Therefore, the product of F_g and Δy is negative. By convention, the work done on the weight should be positive and therefore the quantity $F_g \Delta y$ needs to be given a negative sign to fulfill these requirements. The analysis can then proceed to give the result that the gravitational force is defined by the gradient of the gravitational potential energy:

$$F_{\mathrm{g}} = -\frac{dU}{dy} \tag{22.11}$$

Both of these cases indicate that a force can be described that causes a flux of a material down a gradient. Interestingly, in the case of diffusion, there is actually no directed force like that of gravity or an electric force field that acts on the particles of a species. As will be seen shortly, there is instead an uneven distribution of the particles with respect to position. This phenomenon can be formally treated as if it were a directed force, but in fact it is only a *pseudoforce*. This diffusional "force" will be shown to be closely related to the entropy of mixing.

The experimental description of the net transport of materials by diffusion is quite simple. Again, considering the rectangular volume described earlier, the amount of substance i that travels through a yz-plane of unit area (1 cm^2, for example) in a given amount of time (usually 1 s) is called the *flux* of species i. Flux in this example would be given in units of mol/cm^2/s. The concentration gradient will determine the flux through the yz-plane. If there is no concentration gradient, then there will be no flux, and, conversely, when the concentration gradient is steeper, the measured flux will be higher. *Fick's first law* describes these experimental relationships:

$$J_i = -D\,\frac{dc}{dx} \tag{22.12}$$

D is a proportionality constant called the *diffusion coefficient*. Fick's first law indicates that there is a decrease in the number of particles proportional to the concentration gradient; that is, the transport of species is in a direction opposite to the concentration gradient. The concentration gradient will have units of mol/m^4 and the diffusion coefficient is written as m^2/s. Clearly, the flux of a species i will depend on the apparent force causing the net transport. This force was described in Eq. (22.8) as

$$F_{\mathrm{D}} = -\frac{d\mu_i}{dx} \tag{22.8}$$

This equation can be combined with the general relationship described earlier for transport phenomena under the conditions of a sufficiently small driving force such that the flux is linearly related to the driving force:

$$J_i = -B\,\frac{\partial A}{\partial x} = -BF_A \tag{22.1}$$

The driving force that will cause the transport across the transit plane where flux is being measured will be given by the quantity $c_i\ d\mu_i/dx$ where c_i is the concentration of species i in the plane adjacent to the transit plane. Therefore, Eq. (22.1) can be written:

$$J_i = -Bc_i \frac{d\mu_i}{dx} \tag{22.13}$$

Since $\mu_i = \mu_i{}^{o} + RT \ln c_i$, Eq. (22.13) can be rewritten entirely in terms of concentration as

$$J_i = -Bc_i \frac{RT}{c_i}\frac{dc_i}{dx} = -BRT\frac{dc_i}{dx} \tag{22.14}$$

and

$$D = BRT \tag{22.15}$$

This gives the relationship between Fick's law and the phenomenological theoretical treatment of a transport process under nonequilibrium conditions.

Is the diffusion coefficient, D, constant with respect to concentration? Experimentally, it can be shown that D varies with concentration. The reason for this should be obvious if the assumptions of this analysis are considered. The solution under consideration was assumed to be ideal and this assumption led to Eq. (22.14). We know that solutions of biological importance are not ideal and that their activities vary considerably with concentration. Taking this into account gives the following result:

$$J_i = -Bc_i \frac{d\mu_i}{dx} \tag{22.13}$$

Now explicitly taking into account the activity of i, $\mu_i = \mu_i^o + RT \ln \gamma_i c_i$ leads to

$$J_i = -Bc_i \frac{d}{dx}\left(\mu_i^o + RT \ln \gamma_i c_i\right) \tag{22.16}$$

Solving the equation gives the following result:

$$J_i = -BRT\left(1 + \frac{d\ \ln \gamma_i}{d\ \ln c_i}\right) \tag{22.17}$$

Now rewriting Eq. (22.15) in terms of Eq. (22.17) gives

$$D = BRT\left(1 + \frac{d\ \ln \gamma_i}{d\ \ln c_i}\right) \tag{22.18}$$

The diffusion coefficient does depend on concentration though the dependency is only significant when the activity coefficient varies in the range of concentrations causing diffusion in the first place.

Fick's first law applies to diffusion events where the rate of transport is steady state; that is, even though there is net flux in the system, the rate of flux is constant and the driving force, i.e., –*dc*/*dx*, remains constant. In physical terms, this means that the movement of species *i* through the transit plane of the rectangular volume is constant and that at any time the same amount of material is leaving each *yz*-plane as is entering that plane. If the flux through a small volume, with the dimensions of a *yz*-plane by an *x* + *dx* coordinate, is such that less material is leaving one side of the volume than is coming into the volume from the other side, then the concentration in the volume is increasing with respect to time. Hence, the following partial derivative can be written as

$$\left(\frac{\partial c}{\partial t}\right)_x = -\left(\frac{\partial J}{\partial x}\right)_t \tag{22.19}$$

Using Fick's first law, the change in concentration in a small volume that varies with respect to time can be found as

$$\left(\frac{\partial c}{\partial t}\right)_x = D\left(\frac{\partial^2 c}{\partial x^2}\right)_t \tag{22.20}$$

This is *Fick's second law* and is accurate if the diffusion coefficient is independent of the concentration. Fick's second law is a second-order partial differential equation and the solution of such equations can be difficult. Its solution will not be covered here. Fick's second law is the basis for the treatment of many non-steady state or time-dependent transport problems (Fig. 22.2). Experimental use of Fick's second law is often made to determine diffusion coefficients. It is important to be sure that the experimental conditions and analysis of the data take into account the dependency of the diffusion coefficient on the concentration of the measurement system.

22.3 The Random Walk Forms the Basis for a Molecular Picture of Flux

What is a reasonable molecular picture of the movement of the particles making up a solution at equilibrium? A clue can be gained by observations on the motions of small colloidal particles of dye. These particles, if observed under a microscope, show a haphazard irregular motion of each particle that ultimately leads nowhere. The behavior of each colloidal particle, and in fact of each molecule in the solution, can be qualitatively described and depicted as shown in Fig. 22.3. The irregular zigzagging path of the particle is due to the multiple collisions of that particle with other particles in the solution. The average distance a particle will travel between

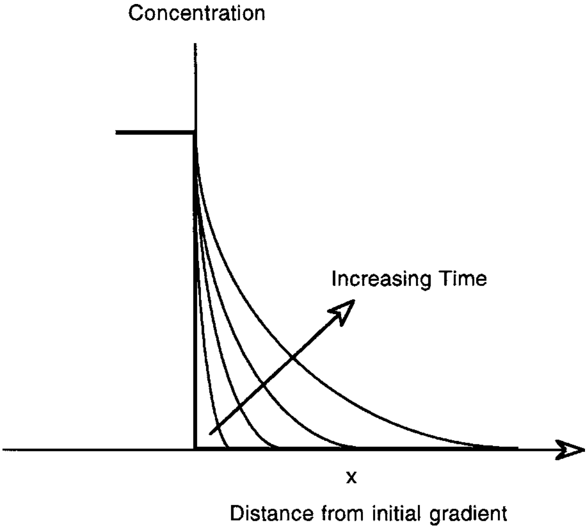


Fig. 22.2 Fick's second law expresses the change in concentration at a point *x* with respect to time, generating concentration profiles like those shown here. The relationship can be used to determine diffusion coefficients experimentally. The concentration profiles drawn in this figure are representative of an experimental electrochemical system in which a new species (R: O + ne^- $\rightarrow$ R) is produced at a diffusion-controlled rate. The current flow in the system will depend on these profiles and the current-time transient contains the diffusion coefficient as a constant of proportionality

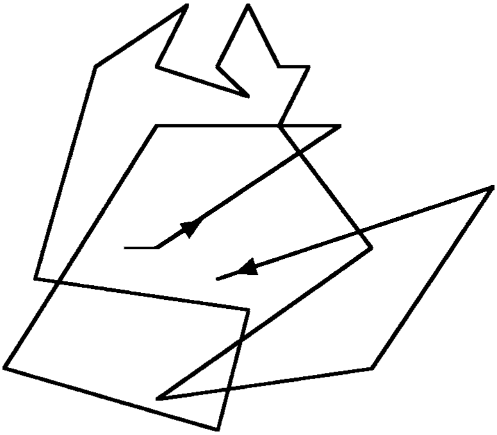


Fig. 22.3 Random wandering of a particle in a solution at equilibrium

collisions is called the *mean free path*. The mean free path, *l*, is given by dividing the average distance a particle will travel in unit time if unimpeded, $\langle u \rangle$, by the number of intermolecular collisions the particle will experience in that time, *z*:

$$l = \frac{\langle u \rangle}{z} \tag{22.21}$$

One measure of the average velocity of molecules is the *root-mean-square velocity*, which we derived in Chapter 7 for an ideal gas or solution:

$$\left\langle u^2 \right\rangle^{1/2} = \left(\frac{3RT}{M} \right)^{1/2} \tag{22.22}$$

The mean velocity of a molecule that includes consideration of the Maxwell–Boltzmann distribution of velocities results by integrating over a continuous distribution of velocities. The Maxwell–Boltzmann distribution is

$$F(u)du = 4\pi \left(\frac{M}{2\pi kT} \right)^{3/2} u^2 e^{-mu^2/2kT} du \tag{22.23}$$

with u the magnitude of the velocity, m the mass of the particle, and the value of interest in the interval du. When the average velocity u^2 is integrated

$$\left\langle u^2 \right\rangle = \int_0^\infty u^2 F(u) du \tag{22.24}$$

using standard tables of integration the result is

$$\left\langle u^2 \right\rangle = \frac{3kT}{m} \tag{22.25}$$

where $\frac{R}{M}$ can be written in Eq. (22.25) to replace $\frac{k}{M}$ with the resulting relationship to Eq. (22.22) is self-evident. To calculate <u> a similar integration is performed:

$$\langle u \rangle = \int_0^\infty uF(u) du \tag{22.26}$$

with the result

$$\langle u \rangle = \left(\frac{8RT}{\pi M} \right)^{1/2} \tag{22.27}$$

where M is the molecular weight.

The number of collisions, z, depends on the concentration of the particles $\frac{N}{V}$, their diameter, σ, and the mean velocity, written as

$$z = 4\sqrt{\pi} \frac{N}{V} \sigma^2 \left(\frac{RT}{M} \right)^{1/2} \tag{22.28}$$

Figure 22.3 suggests that although the net distance a particle may travel from its origin will be zero, the total distance traveled by the particle in a particular time, t, can be significant. Knowledge of the distance traveled is necessary for a molecular understanding of the events of diffusion. In a single dimension, the average distance from the origin will be determined by summing the distance covered by all of the steps the particle takes in a positive or negative direction from the origin in a large number of tries, *i*, and then dividing by the number of tries:

$$\langle x \rangle = \frac{\sum_i x(i)}{\sum i} \tag{22.29}$$

Obviously, for a large enough number of tries, the number of moves a particle makes away from the origin to the left will be equal to the number of moves to the right and $<x>$ will equal zero. This is the mean progress. This result is the consequence of adding the positive and negative forays from the origin together. If each of the jumps is squared and then summed, a value that is positive will result. This is the *mean square distance*, $<x^2>$. Since the distance that is traveled with each step or start is the same (i.e., the mean free path of the particle) and the number of steps is increasing at a constant rate with time, the mean square distance traveled increases in a linear fashion with time; that is, $<x^2>$ is proportional to time.

Now, it is possible to point out the relationship between the random-walking particles and diffusion. Recalling the rectangular box used earlier, consider the random walk of some solute molecules, for example, ions in solution. A *yz*-plane is dropped at the origin of the *x*-axis. Two more *yz*-planes are now established on either side of the origin at $-\left(<x^2>\right)^{1/2}$ and $+\left(<x^2>\right)^{1/2}$. Therefore, two equal volumes, *a* and *b*, now exist on each side of the plane at the origin, each with a concentration of the ion species *i*, c_a and c_b (see Fig. 22.4). Both volumes, *a* and *b*, start with the same number of molecules of *i*. The ions move in a random walk along the *x*-axis. In a particular time, *t*, all of the ions in compartment *a* that are moving in the left-to-right direction will cross the *yz*-plane at the origin (this is ensured by choosing the dimension $<x^2>^{1/2}$ for the volume). Since there is an equal probability of an ion going left to right or right to left across the plane at the origin, the flux across the transit

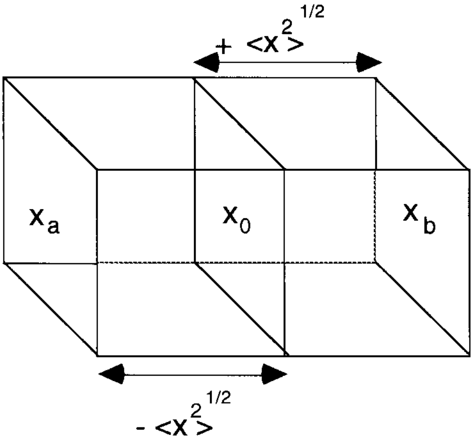


Fig. 22.4 System to study the random walk approach to diffusional forces

plane from *a* to *b* will be given by 1/2 $<x^2>^{1/2}c_i$. An equivalent analysis of the flux across the transit plane from *b* to *a* via a random walk can be made. If the number of ions in each volume is equal at the end of *t*, there will be no net flux, which is the result that is expected. The flux from side *a* to *b* can be written as follows:

$$J = \frac{1}{2}\frac{\left\langle x^2\right\rangle^{1/2}}{t}\left(c_a - c_b\right) \tag{22.30}$$

If the number of ions of species *i* found in the two volume elements, *a* and *b*, is not the same, there will be a measurable net flux simply because of the random walk mechanism. On a molecular basis then, it is clear that diffusion occurs without a specific diffusive force that acts to move each particle. As was mentioned previously, this is why diffusion can be formally treated as a force but in fact is different in nature from an electrostatic or gravitational force. The relationship of the random walk-based molecular model to Fick's laws can be shown as follows. The concentration gradient from side *a* to *b* is evaluated:

$$\frac{\partial c}{\partial x} = \frac{c_b - c_a}{\left\langle x^2\right\rangle^{1/2}} = -\frac{c_b - c_a}{\left\langle x^2\right\rangle^{1/2}} \tag{22.31}$$

Writing this result in terms of the difference ($c_a{-}c_b$) gives

$$\left(c_a - c_b\right) = -\left(\left\langle x^2\right\rangle\right)^{1/2}\frac{\partial c}{\partial x} \tag{22.32}$$

This result can be directly substituted into Eq. (22.30), which after simplifying gives

$$J = -\frac{1}{2}\left(\frac{\left\langle x^2\right\rangle}{t}\right)\left(\frac{\partial c}{\partial x}\right) \tag{22.33}$$

Equating Eq. (22.33) with Fick's first law gives

$$-\frac{1}{2}\left(\frac{\left\langle x^2\right\rangle}{t}\right)\left(\frac{\partial c}{\partial x}\right) = -D\,\frac{dc}{dx} \tag{22.34}$$

Simplifying yields

$$\frac{\left\langle x^2\right\rangle}{2t} = D \tag{22.35}$$

Rearrangement gives the *Einstein–Smoluchowski equation*:

$$\left\langle x^2\right\rangle = 2Dt \tag{22.36}$$

The factor 2 in this equation is a result of the one-dimensional derivation. A three-dimensional derivation will have a coefficient of 6.

Diffusion processes in biological systems often occur between phases of significantly different structure and composition, such as from the generally aqueous environment of the blood through the bilipid layer of the cell membrane. The relationship of the diffusing species to the solvent clearly must be considered. The diffusion coefficient has been treated as a phenomenological proportionality coefficient, but clearly, if considered on a molecular basis, it contains information about solute–solvent interactions. If the random-walking ion is intuitively considered, the mean square distance would be expected to depend on the distance moved with each jump and the number of jumps that occur in a unit of time, *t*. These values have been discussed already – the mean free path, *l*, and the number of collisions, *z*. It can be shown that

$$\left\langle x^2 \right\rangle = zl^2 \tag{22.37}$$

Combining Eqs. (22.36) and (22.37) yields the result for a one-dimensional random-walking ion:

$$zl^2 = 2Dt \tag{22.38}$$

In the case where $z = 1$, that is, only a single jump is considered, this expression reduces to

$$D = \frac{l^2}{2\tau} \tag{22.39}$$

where τ is equal to the mean jump time for the particle under consideration to cover the mean jump distance, *l*. τ is the period of the cycle of the jump including time between the jumps. The frequency of the jump is simply $1/\tau$. The diffusion coefficient then is dependent on the mean jump distance and the frequency of the jumps.

What is the physical consequence of Eq. (22.39)? First of all, the distance that a particle will be able to jump is obviously limited by the physical structure of the solvent in which it is jumping. Generally, this means that the particle is going to jump from an occupied site in the solvent to an unoccupied site in the solvent. After the jump, there will have been an exchange of occupied and unoccupied sites. The mean free path or mean jump distance will be dependent on the structure of the solvent at the instant of the jump.

While the mean free path is essentially dependent on the solvent structure, the frequency of the jumps is a question of rate. Application of rate process theory, in which the free energy change from the pre-jump to post-jump coordinate goes through an activated state of higher free energy, provides an approach to this problem. Since the particle must cross an energy barrier to go from site to site, the frequency of the jumps will be given by a rate constant, k_{forward}:

$$k_{\text{forward}} = \left(\frac{kT}{h}\right) e^{-\Delta G^{\dagger}/RT} \tag{22.40}$$

where k is Boltzmann's constant, h is Planck's constant, and $\Delta G^{\dagger}$ is the free energy of the transition state.

The diffusion coefficient can therefore be considered in terms of the mean jump distance and the frequency of the jumps. Because all jumps are of interest in a problem of diffusion, the coefficient in the denominator from the Einstein–Smoluchowski equation can be taken as unity, giving

$$D = \left(\frac{l^2 kT}{h}\right) e^{-\Delta G^{\dagger}/RT} \tag{22.41}$$

Both molecular and empirical sketches of the transport phenomenon of diffusion have been presented so far. The driving force that causes movement down a chemical potential gradient was shown to be a pseudoforce. Next, the movement of particles in an electric field will be discussed. While the nature of the forces in these two cases is significantly different, it will be found that the formal treatment of these forces is quite similar.

Further Reading

For a more detailed presentation of diffusion including the use of Fick's second law see

Berg, H.C. (1993) *Random Walks in Biology*, Revised edition. Princeton University Press, Princeton, NJ.

Bockris J.O'M, Reddy A.K.N., and Gamboa-Aldeco M. (1998) *Modern Electrochemistry*, 2nd edition, Volume 1. Kluwer Academic/Plenum, New York.

Dill K.A. and Bromberg S. (2003) *Molecular Driving Forces: Statistical Thermodynamics in Chemistry and Biology.* Garland Science, New York.

Other Articles of Interest

Dix J.A. and Verkman A.S. (2008) Crowding effects on diffusion in solutions and cells, *Annu Rev. Biophys*., **37**:247–263.

Einstein A. (1956) *Investigations on the Theory of the Brownian Movement.* Dover Publications, New York.

Problem Sets

1. Explain in qualitative but succinct terms the following statement: "Entropy is the primary driving force in transport phenomena." What does this imply about the timescale of transport phenomena?

2. Compared to an aqueous solution at 298 K, would you expect the mean free path of a sodium ion to be longer or shorter in

   (a) Ice at 273 K.
   (b) Aqueous solution at 273 K.
   (c) Aqueous solution at 373 K.

Chapter 23
Flow in an Electric Field: Conduction

Contents

23.1 Transport of Charge Occurs in an Electric Field

To this point, our discussion has focused on the transport of molecules simply by the random process of diffusion. What are the events when the system experiences an electric field? This is an extremely important aspect of cellular systems, because so many biologically active chemical species, including proteins, carbohydrates, lipids, and ions, carry a net charge and will respond with a net flow of current to the electric fields that are ever present in cellular systems. In the following derivation, the focus will be on the net movement of charge in solutions of inorganic ions, because the majority of charge transfer in solution is accounted for by these ions.

An electrolyte solution has been shown to contain a certain amount of charge, given by its ionic strength, but balanced precisely so that the condition of electroneutrality is met. Under the conditions of only thermal randomizing forces (i.e., no chemical, gravitational, electric, or other force fields), the ions and solvent molecules are in constant motion without any net movement or transport. When an electric field is imposed across this system there will be a conduction of electric current by the electrolyte solution. The conductivity of an electrolyte solution is primarily due to the addition of the ionic species. The conduction of electricity in

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_23, © Springer Science+Business Media, LLC 2010

solution is due to the net transport of ions. We proceed by first drawing an empirical sketch of the process and then provide a more detailed molecular description of the process.

23.1.1 Ionic Species Can Be Classified as True or Potential Electrolytes

If ions can behave as carriers of electric charge, then they must be free to move in the presence of an electric field. How can a system that contains mobile ions be produced? There are several possible methods for producing these mobile ions. Each method reflects the chemical nature of the parent electrolyte. The term electrolyte, as used in biomedicine and electrochemistry, refers to both the ionically conducting medium and the substances that produce electrolytes when they are dissolved or liquefied. A specialized form of electrolyte is the colloidal electrolyte, a compound that gives rise to at least one ionic macromolecular species that becomes the ionic conductor.

Materials, such as salts (e.g., sodium chloride), will conduct a current in the solid state whether or not they are in solution. It would be a reasonable inference that such materials must be comprised of ions in their pure state. Such electrolytes are named ionophores or true electrolytes. A common nomenclature also calls such electrolytes as strong electrolytes because when dissolved in water they conduct strongly. This is in contrast to substances called weak electrolytes, which do not give rise to a high conductance when dissolved in water (see the following paragraph). This historical nomenclature can lead to some problems in nonaqueous systems. In this book, the term *true electrolyte* will be used. True electrolytes in the solid form are comprised of a lattice of positive and negative ions arranged in highly ordered and repeating structures. If a crystalline true electrolyte is heated, the translational motion of the ions will increase. A point is reached where the lattice structure is disrupted and the crystal melts. This liquid salt is a liquid ionic conductor, because the ions carrying the charge are now free to move in a liquid phase. A true electrolyte therefore is not required to dissolve in a solvent to become a liquid ionic conductor. However, the temperatures necessary to generate a liquid ionic conductor from a pure salt are significant (500–1300 K) and certainly not compatible with life on this planet. When a true electrolyte is dissolved in an aqueous solvent, the ions are displaced from their lattice and become associated with the solvent molecules. A liquid conducting solution is generated that does not require the addition of heat to dissociate the ions from their lattice. This event leads to questions about the interactions between the ions and the water molecules that allow the ions to leave the ion lattice structure and carry charge in the aqueous solvent. Indeed, this is the intuitive process that should previously have suggested that there will be solute–solvent interactions in ionic solutions, at least those composed of true electrolytes, such as salts.

When a similar analysis is applied to certain other materials, for example, an organic acid, such as butyric acid, it will be found that in the pure state there is little electrical conduction. This lack of conduction occurs because the carboxylic acids are not ionized in their pure forms. Without ions, no charge transfer can occur. There is a significant difference between the liquid form of a true electrolyte and the liquid form of these organic acids. When an organic acid such as butyric acid is dissolved in water, however, an interesting change in behavior takes place; the solution of the acid in water becomes a liquid ionic conductor. By definition, the butyric acid is an electrolyte since it gives rise to a liquid ionic conductor, but, clearly, a different set of events has taken place in this scenario compared to the case of the true electrolyte. The difference is that butyric acid and other carboxylic acids in solution with water will undergo dissociation into an acid–base pair, both of which are ionized and therefore will carry an electric charge. Electrolytes such as these, whose ionic character depends on a chemical reaction to generate their ionic nature, are called potential electrolytes. It is evident that a potential electrolyte must interact with its solvent, since it depends on this solvent–solute interaction to become a conducting solution. Historically, electrolytes that gave rise to limited conductance when dissolved in water were called weak electrolytes. This is a solvent-specific classification (a “weak” electrolyte in water may give rise to “strong” electrolyte conducting properties in another solvent), and the preferred term of *potential electrolyte* is used in this book.

23.2 Describing a System of Ionic Conduction Includes Electronic, Electrodic, and Ionic Elements

An external electric field is imposed on the electrolyte solution and the ions experience an electric force superimposed on the thermal randomizing forces that lead to the random walk behavior of the molecules. To understand the effects of the external electrical force on the ions in solution, it is necessary to have a description of the electric field that is created in the electrolyte solution upon the application of an external field. The system to be placed under study will be similar to the rectangular box used earlier for the examination of diffusion except that now a pair of parallel plate electrodes are placed at the ends of the box. These electrodes are connected through a circuit to an external power source. Figure 23.1 illustrates this system.

An electrolyte solution fills the box between the electrodes and a switch connects the battery to the circuit. Charge in the form of electrons flows from the negative terminal of the battery through the wires of the external circuit into the negative electrode or *cathode*. At the cathode–electrolyte interface, a charge transfer reaction occurs in which the electron, which is responsible for carrying charge in the external circuit, is transferred via an *electronation* or *reduction reaction* to an ion in the electrolyte phase. In the electrolyte phase, the ions carry the charge down the potential gradient to the positive electrode, the *anode*. At the anode, the charge carried by the ion will be transferred to the anode via an *oxidation* or *de-electronation reaction* and will complete the trip around the external circuit as an electron. This description

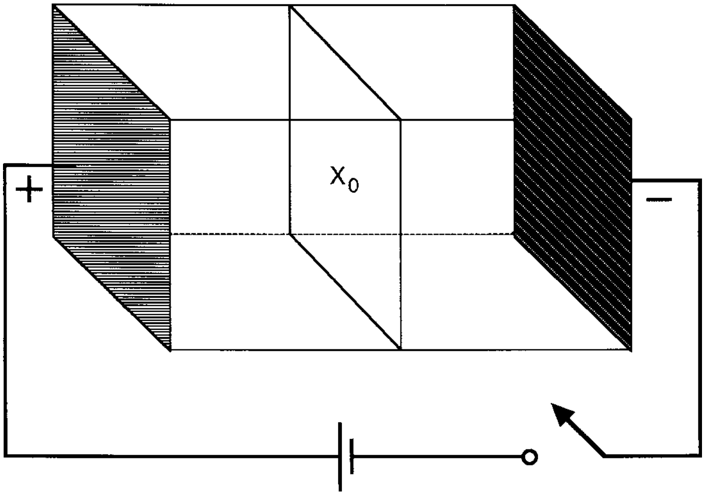


Fig. 23.1 Experimental system for the description of electrical conduction

of charge transfer in a circuit comprised of both ionic and electronic charge carriers leads to the classifications used by electrochemists to discuss the system. The study of the conduction and behavior of the electrons in the external circuit is called *electronics* while the study of the charge transfer by ions in solution is called *ionics*. The extremely interesting and important study of the charge transfer reactions at the electrodes is called *electrodics* and is the focus of modern electrochemistry. The focus of this chapter will be on ionics, but the study of electrodics as related to biological systems is an important field in bioelectrochemical research. The basic principles of bio-electrodics are discussed in the following chapters.

If conduction is taking place in the electrolyte solution, a driving force or potential gradient field must be present down which the ions move. A potential map of the system in Fig. 23.1 can be drawn by finding the work necessary to bring a test charge from infinity to any point x in the system. This work will define the potential, ψ, at each point. Such a potential map is illustrated in Fig. 23.2. It is notable that the largest potential drops in the system occur at the two electrode–electrolyte interfaces. This is not surprising given the events of electronation and de-electronation occurring at the electrodes. The work done to bring a test charge through the interface involves transfer between two phases and also the generation or consumption of chemical species. Our focus here is on the potential gradient between the two electrodes. This is the gradient that the ions in the electrolyte solution are experiencing. By choosing a design with parallel plate electrodes, the electric field, X, will be linear with respect to distance as Fig. 23.2 shows. The strength of the electric field is found by taking two equipotential planes such as x_0 and any other plane x_j, finding the potential difference between them, and dividing by the distance separating them, d:

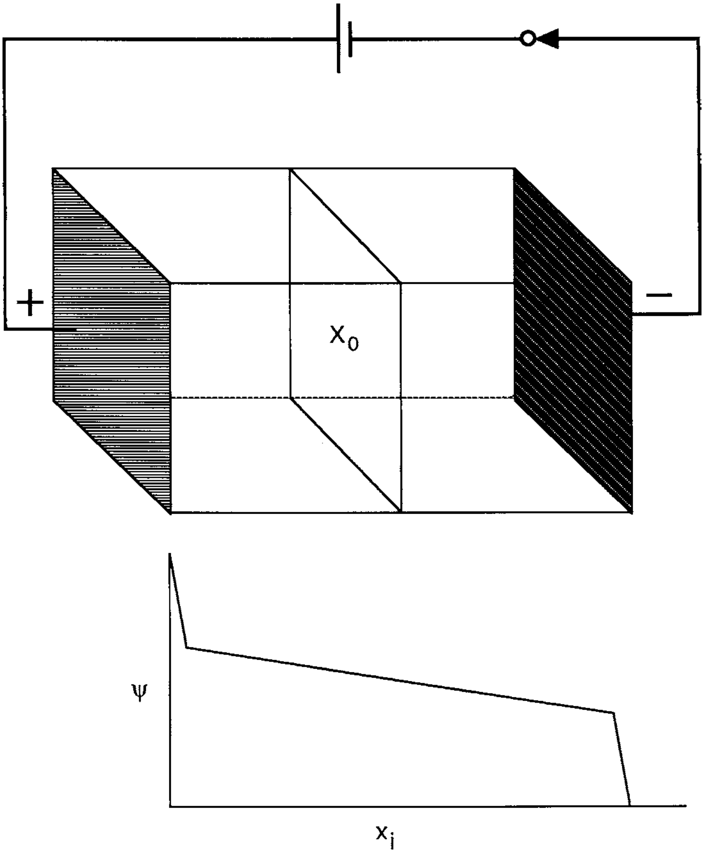


Fig. 23.2 Voltage drops in the ionic and electrodic portions of the experimental circuit

$$E = -\frac{\Delta\psi}{d} \tag{23.1}$$

The negative sign occurs because E is the negative of the gradient of the potential. A series of equipotential planes parallel to the electrodes will exist for this experimental system, providing a linear gradient in a fashion similar to the example developed for diffusion. Although in this case the gradient of the field responsible for transport is linear, this does not have to be the case and the potential field can generally be written as

$$E = -\frac{d\psi}{dx} \tag{23.2}$$

We have already described the picture of molecular movement in a solution. Consider the experimental circuit of Fig. 23.1 before the switch is thrown. We focus on the movement of a single ion in an electrolyte solution. The ion is found in the quasi-crystalline water vibrating about an equilibrium point with no net jumping. At a particular time, the ion will have attained sufficient energy to jump to a free or unoccupied site elsewhere in the water. The rate of the jumping is on the order of 10^{10} to 10^{11} s^{-1} for aqueous systems of biological interest. The jumps occur with a frequency that depends on the energy needed to separate the ion from its first site and on the energy needed to form a hole into which the jumping ion can go. The electric field in the solution is now switched on and the ions experience an electrical force. The positive ions are attracted to the negative electrode (cations) and negative ions are attracted to the positive electrode (anions). This added force serves to nudge the jumping ions in the preferred direction of the appropriate electrode, leading to an overall net flux of current. This flux is the ionic conduction. Figure 23.3 compares the motion from the random walk to the motion of an ion experiencing an electric field. The velocity of motion imparted to the ions by the electric field is small compared to the instantaneous velocity resulting from the thermal collisions. For example, the ionic velocity for a 1 V/m field will be approximately 50 nm/s, while the thermally induced instantaneous velocity can be 100 m s^{-1}. The molecular picture of ions in an electric field is like that of a crowded lobby during intermission at the theater. Before the bell rings to summon the crowd for the second act, people are milling about with constant motion, but no actual net movement into the theater itself occurs. At the call for the second act, people begin to move toward their seats, though certainly not in a direct manner. Ultimately, there is a net flux from the lobby to the house and each individual may be tracked zigzagging, starting, and stopping along the way.

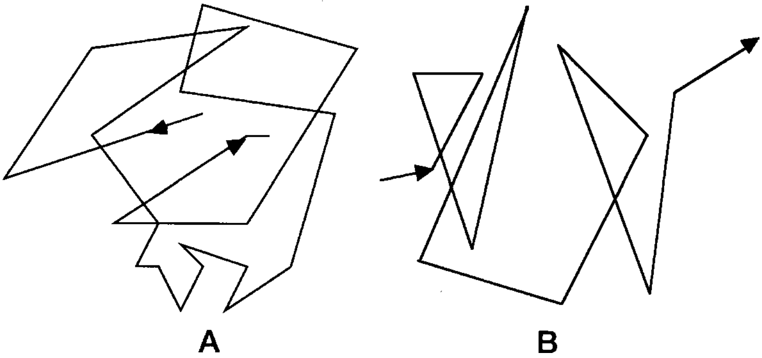


Fig. 23.3 Comparison of the path of an ion on a random walk (**a**) versus that of an ion in an electric field (**b**)

Conduction of charge will continue as long as the net electroneutrality of the solution is maintained. The charge added to the solution at one electrode must be removed at the same rate at the opposite electrode. If this were not the case, the electrolyte would almost immediately attain a net charge and, as in the case of the capacitor discussed previously, a counter field would rapidly develop that would prevent the addition of any further charge to the solution; the flow of current would cease. The reactions at the electrodes therefore are the vital link to the continued flow of charge in these systems.

23.3 The Flow of Ions Down a Electrical Gradient Can Be Described Phenomenologically

After the electric field has been applied to the electrolyte solution for some time, a steady-state flow of ions will be observed. The steady-state flow depends on the preservation of electroneutrality in the solution by charge transfer reactions at the two electrodes and *Faraday's law* expresses the quantitative relationship for the transfer of electronic charge to ionic charge that occurs at the electrodes:

$$\frac{m}{FW} = \frac{it}{|z_i|\,F} \tag{23.3}$$

where m is the mass of an element of formula weight FW liberated or consumed at an electrode, i is the current passed in amperes in t seconds, z_i is the charge on a given ion, and F is the Faraday, equal to 96,484.6 C/mol. When an electrical potential field is imposed across a conducting material, the amount of current flow depends on both the potential and the resistance to flow of current, the *resistance*. This is the relationship quantified in *Ohm's law*.

While electrolyte solutions obey Ohm's law (at least for small potential fields and steady-state conditions), the current flux depends on the concentration of the ions making up the solution, as well as their identity and charge. Consequently, the conductivity of an electrolyte solution varies with concentration (see Fig. 23.4). Because of the variation in conductivity with concentration, it is important to determine the conductance on a per particle basis and to introduce the concept of *molar conductivity*, Λ_m:

$$\Lambda_\mathrm{m} = \frac{\kappa}{c} \tag{23.4}$$

The units of Λ_m are Ω^{-1} cm^2 mol^{-1} or Ω^{-1} m^2 mol^{-1}. The molar conductivity is valuable when comparing the conductivity of solutions that may not have the same concentration. The comparison of molar conductivity, however, is complicated by the fact that different types of electrolytes, i.e., 1:1, 1:2, will contain anions and cations of different charge. To normalize for this variable, the *equivalent conductivity*, Λ, can be used:

$$\Lambda = \frac{\kappa}{cz_i} \tag{23.5}$$

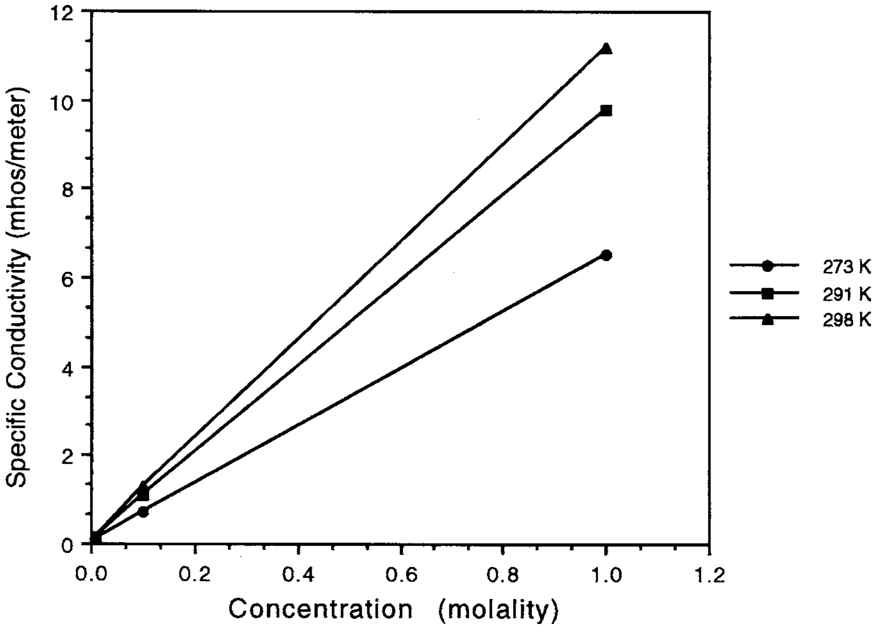


Fig. 23.4 Conductivity,κ, of KCl solutions at varying concentrations and temperatures

What is apparent when the molar or equivalent conductivities are graphed against the concentration of the solution is that the Λ actually falls as the concentration is increased. In Fig. 23.5, the data from the graphs in Fig. 23.4 are expressed as Λ rather than as conductivity, κ, of the solution. This result may be surprising, since it would seem at first glance that the increase in potential charge carriers should lead to an increase in the flow of charge in the conductor, as is shown in Fig. 23.4. However, the equivalent conductivity data indicate that the increase in the concentration of charge carriers leads to a diminished ability of each ion to transport charge. The shape of the Λ versus concentration curve suggests that the maximum conductivity for each ion occurs in the limit of an infinitely dilute solution. This observation invites two conclusions: If the maximum conductivity per ion occurs under the conditions of an ideal solution, then the molecular reason for the observed decrease in equivalent conductivity will probably lie in interactions between the ion and other ions, as well as the ion and its solvent, as is seen in the case of the activity coefficient. It also suggests that if a series of experimental points are obtained, and an extrapolation is made to the infinitely dilute solution, a reference state of maximum equivalent conductivity can be defined that will be useful for comparisons between any electrolytes. In Fig. 23.6, the method of determining this reference state, Λ^0, the *equivalent conductivity at infinite dilution*, is illustrated. Table 23.1 lists the equivalent conductivity at infinite dilution for a variety of electrolytes.

The dependence of the equivalent conductivity of true electrolytes such as KCl on the concentration was determined by Kohlrausch, who showed that Λ varied with

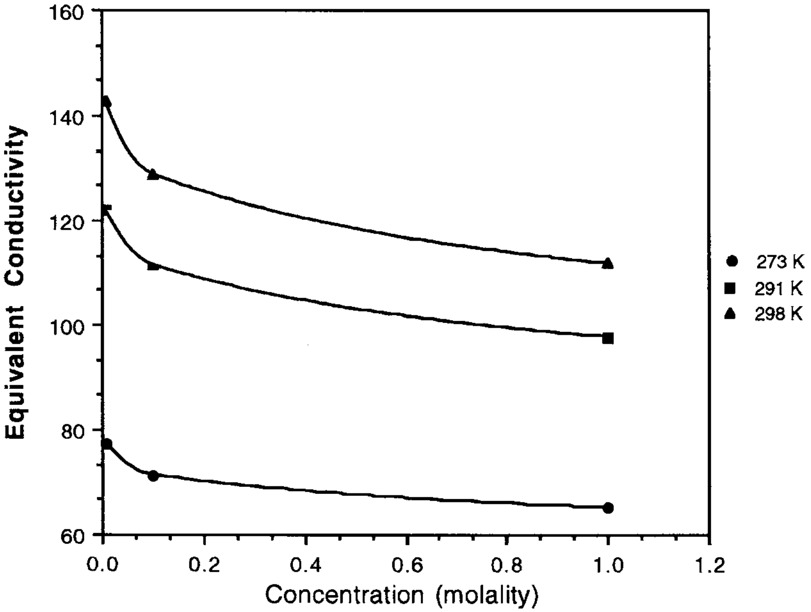


Fig. 23.5 Equivalent conductivity, Λ, of solutions of KCl at various concentrations and temperatures

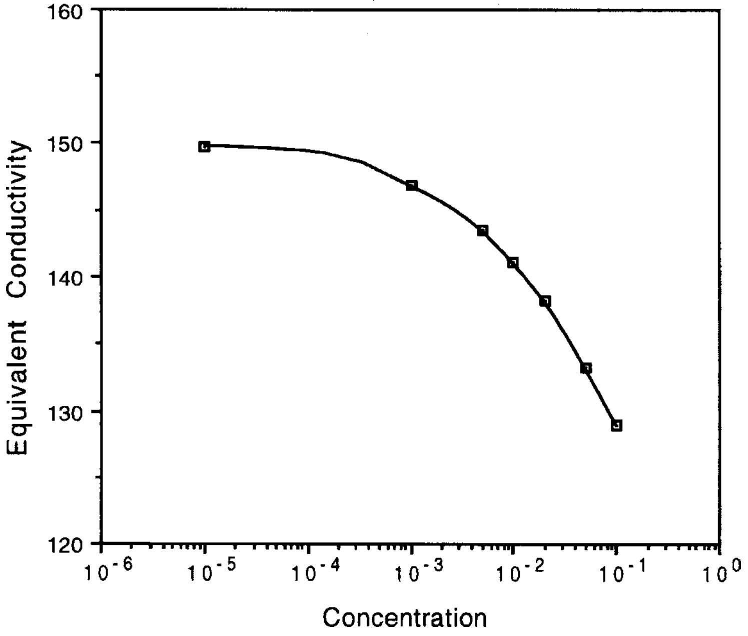


Fig. 23.6 Method for determining the equivalent conductance at infinite dilution, Λ^0, for KCl

Table 23.1 Equivalent conductivity at infinite dilution, Λ^0, for selected electrolyte solutions at 298 K

Electrolyte	Λ^0 (Ω^{-1} m^2 equiv^{-1} × 10^{-4})
HCl	426.16
LiCl	115.03
NaCl	126.45
KCl	149.86
NaOH	247.80
$MgCl_2$	129.40
$CaCl_2$	135.84
$LaCl_3$	145.80

the square root of the concentration at low concentrations:

$$\Lambda = \Lambda^o - kc^{1/2} \tag{23.6}$$

where k is an empirical constant which is found as the positive slope of the straight line obtained by graphing Λ against $\sqrt{c}$ as shown in Fig. 23.7.

Kohlrausch also showed by experiment that the value of Λ^0 for any true electrolyte depends on the sum of the equivalent conductivities at infinite dilution of the anions and cations making up the true electrolyte. These ionic conductivities

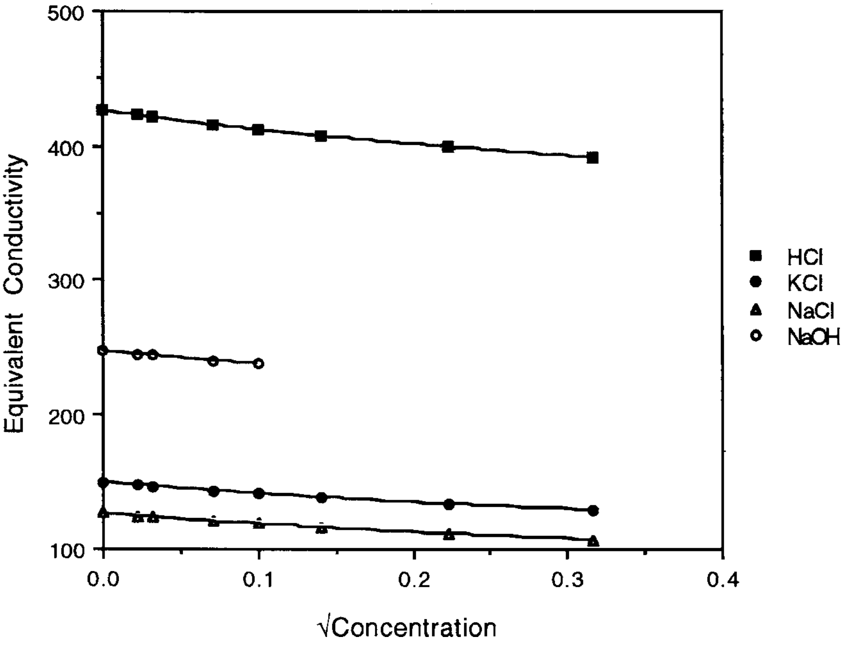


Fig. 23.7 *Straight lines* showing the validity of the Kohlrausch law, indicating the dependence of conductivity on the concentration for true electrolytes

at infinite dilution are given the symbols λ^{o}_{+} and λ^{o}_{-}.This observation is known as Kohlrausch's law of the independent migration of ions:

$$\Lambda^{o} = \nu_{+}\lambda^{\infty}_{+} + \nu_{-}\lambda^{\infty}_{-} \tag{23.7}$$

where ν_{+} and ν_{-} account for the stoichiometry of the cations and anions, respectively, that are carrying the charge; for example, for HCl, $\nu_{+} = 1$ and $\nu_{-} = 1$ while for $CaCl_2$, $\nu_{+} = 1$ and $\nu_{-} = 2$. Equation (23.7) can be used to predict the equivalent conductivity of a true electrolyte at infinite dilution from the equivalent conductivities at infinite dilution of the ions. The values of Λ^{o} in Table 23.1 could be found by using Table 23.2 and Eq. (23.7). This result is consistent with the idea put forth earlier that the fall in equivalent conductance of solutions of electrolytes with increasing concentration is related to the interactions between the moving ions during transport in an electric field, since at infinite dilution it would be expected that no interaction can occur, and the conductivities would be truly independent of other ions. How the experimental evidence presented here fits a molecular model will be the subject of the next section.

Table 23.2 Ionic conductivities at infinite dilution for selected ions at 298 K

Cation	$\lambda^{o}_{+}(\Omega^{-1}\,m^{2} \times 10^{-4}$	Anion	$\lambda^{o}_{-}(\Omega^{-1}\,m^{2} \times 10^{-4}$
H^{+}	349.80	OH^{-}	197.60
Li^{+}	38.69	Cl^{-}	76.34
Na^{+}	50.11	Br^{-}	78.40
K^{+}	73.50	$CH_3CO_2{}^{-}$	40.90
$\frac{1}{2}Mg^{2+}$	53.06		
$\frac{1}{2}Ca^{2+}$	59.50		

Before proceeding to develop a model, it is important to realize that up to this point our discussion has dealt only with true electrolytes. Experiments show significantly different behavior for potential electrolytes. The behavior of potential electrolytes is important because of the prominent role they play in some biological systems. Figure 23.8 shows the dependence of Λ on the concentration for both a true electrolyte, HCl, and a potential electrolyte, CH_3COOH. The potential electrolyte clearly does not behave in a fashion similar to that of the true electrolyte, in that it has a very low equivalent conductivity even at low concentrations, but as the concentration of the electrolyte approaches infinite dilution, the equivalent conductivity increases dramatically toward the levels of true electrolytes. The reason for this behavior is that potential electrolytes are not fully ionized until very low concentrations and therefore the equivalent conductivity is dependent on the equilibrium between the unionized and the ionized forms of the potential electrolyte at higher concentrations.

It would be expected that Kohlrausch's relationship of the dependency of Λ on the $\sqrt{c}$ would not hold in the case of the potential electrolyte and Fig. 23.9 confirms this. The conductivity of a potential electrolyte solution depends on the degree of

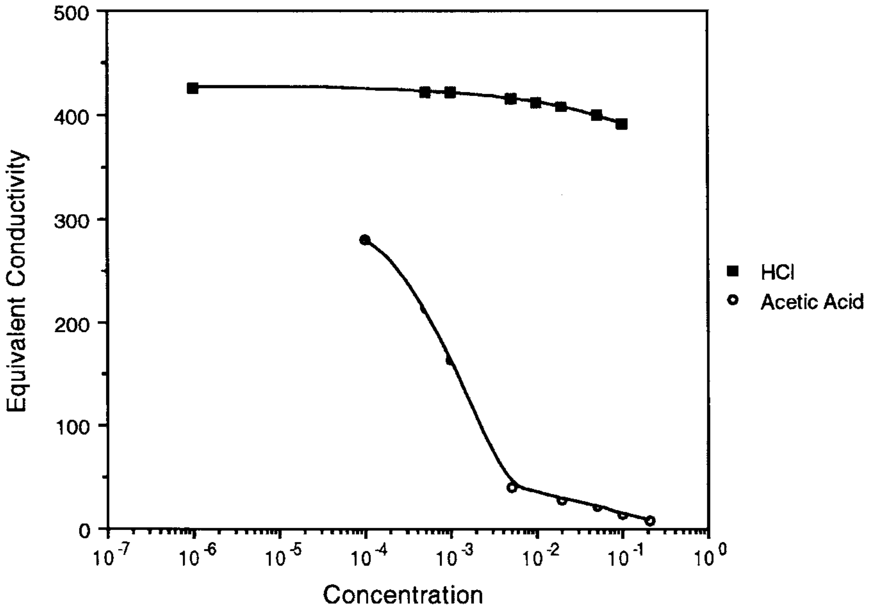


Fig. 23.8 Comparison of the dependence of Λ on concentration for a true and a potential electrolyte solution

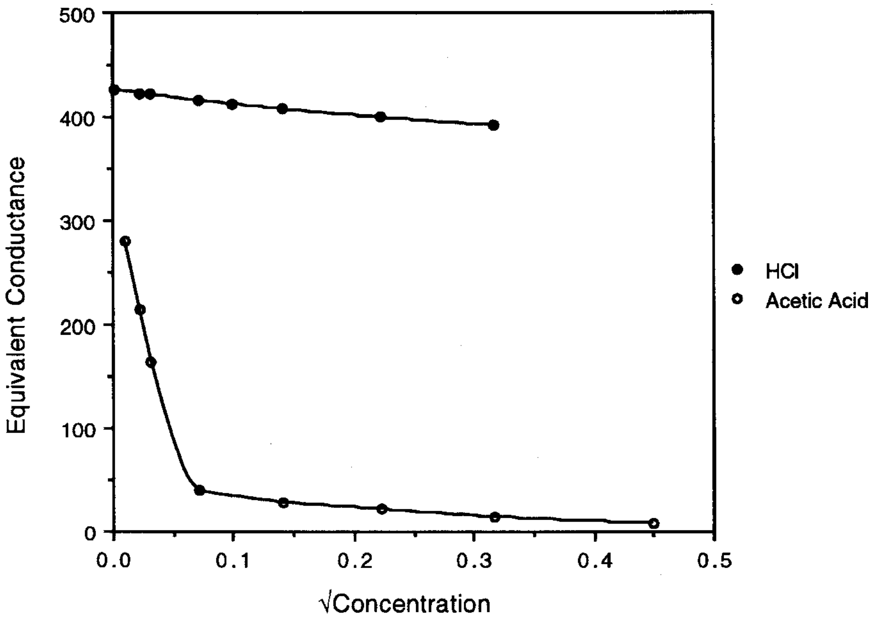


Fig. 23.9 Plotting the equivalent conductance against $\sqrt{c}$ does not give a linear relationship for a potential electrolyte

ionization, α, of the electrolyte. The equilibrium constant of apparent dissociation is related as follows (c is the concentration of the species in the following equations):

$$\begin{aligned} K &= \frac{\left(c_{\mathrm{M}^+}\right)\left(c_{\mathrm{A}^-}\right)}{\left(c_{\mathrm{MA}}\right)} \\ &= \frac{(\alpha c)\,(\alpha c)}{(1-\alpha)\,c} \\ &= \frac{\left(\alpha^2\right)}{(1-\alpha)}c \end{aligned} \tag{23.8}$$

Because the conductivity depends on the degree of ionization, the measured conductivity, Λ', can be related to the equivalent conductivity of a fully ionized solution:

$$\Lambda' = \alpha \Lambda^{\mathrm{o}} \tag{23.9}$$

Kohlrausch’s law of independent migration is true for both potential and true electrolytes, since at infinite dilution α will equal 1. Solving Eq. (23.8) for α and substituting into Eq. (23.9) gives *Ostwald’s dilution law*:

$$K = \frac{\Lambda^2 c}{\Lambda^{\mathrm{o}}\left(\Lambda^{\mathrm{o}} - \Lambda\right)} \tag{23.10}$$

We will now develop a molecular model that can explain the empirical observations described. We will learn that the surprising alterations in ionic conductivity are related to a variety of interactions.

23.4 A Molecular View of Ionic Conduction

The law of independent migration indicates that the total current-carrying capacity of an electrolytic solution depends on the vectorial motion of oppositely charged ions in a solution. Since the current is the sum of all of the charge carriers, this picture is perfectly consistent with the empirical evidence presented so far. A moment’s consideration of Fig. 23.10 leads to several questions: Is it reasonable to assume that except under the most ideal conditions, i.e., infinite dilution, the ions flowing past one another have no contact or interaction? And ignoring this question of interionic interaction, why do some ions have a higher equivalent conductivity (cf. Table 23.2) at infinite dilution when compared to others? These questions must be appropriately answered as we work toward a satisfactory molecular model of ionic conduction.

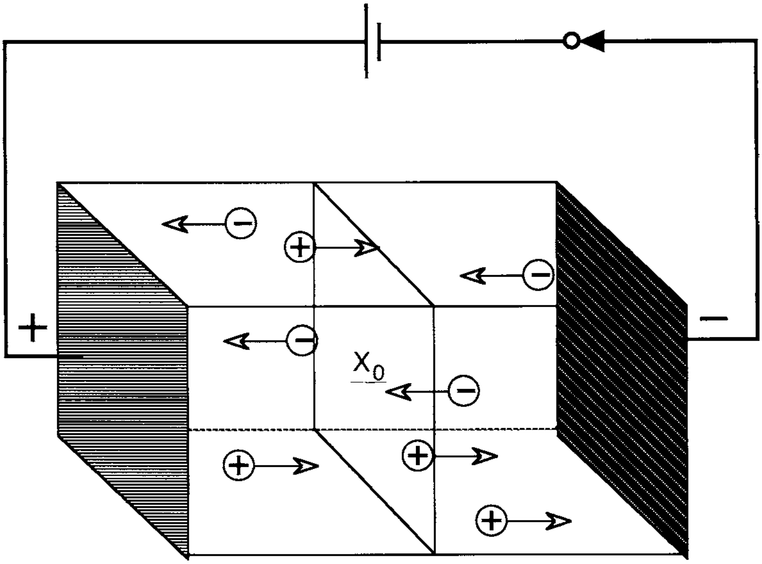


Fig. 23.10 The total current is found by the vectorial sum of all of the charge carriers

First consider a case where the law of independent migration is operating. Any interaction *between ions* that will prevent their independence will be ignored. An ion under an electrical potential field experiences a force given by

$$F = z_i e_0 E \tag{23.11}$$

with the electric field, *E*, described by the negative gradient of the potential:

$$E = -\frac{d\psi}{dx} \tag{23.12}$$

The force on the ion that is generated by the potential field will accelerate the ion up to a maximum velocity, the drift velocity, *v*. The fact that the ion reaches a maximum velocity depends on the drag that acts to retard the movement of the ion. This drag or frictional counter force in a solution is the viscous force. The viscous force is often calculated using *Stokes' law* which for a spherical object is

$$F_v = 6\pi r\eta v \tag{23.13}$$

where r is the radius of the object, η is the viscosity of the solvent, and v is the *drift velocity*. At the drift velocity, the forces acting to accelerate the ion, F_{electric}, and to retard it, F_{viscous}, should be equal:

$$z_i e_0 E = 6\pi r\eta v \tag{23.14}$$

The drift velocity can be written in terms of this balance:

$$v = \frac{z_i e_o E}{6\pi r\eta} \tag{23.15}$$

The drift velocity is related to an index of the ease with which an ion can move, the *mobility*, μ. The absolute mobility is a proportionality constant that defines the drift velocity with respect to a unit force:

$$\mu = \frac{v}{F} \tag{23.16}$$

The mobility of an ion is related to its ability to contribute to carrying a charge in an electric field. Table 23.3 tabulates the relationship.

Table 23.3 Comparison of the ionic mobilities with $\lambda^{\circ}_{\pm}$ in H_2O at 298 K for selected ions.

Ion	μ ($m^2\ s^{-1}\ V^{-1} \times 10^{-8}$)	$\lambda^{\circ}_{\pm}$ ($\Omega^{-1}\ m^2 \times 10^{-4}$)
H^+	36.25	349.82
K^+	7.62	73.52
Na^+	5.19	50.11
Li^+	4.01	38.69
OH^-	20.48	198.50
Br^-	8.13	78.40
Cl^-	7.91	76.34
$CH_3CO_2{}^-$	4.24	40.90

The relationship between conductivity and mobility is of fundamental importance and can be related to current flux through Ohm's and Faraday's laws. These equations are as follows for the individual conductivity:

$$\lambda_{\pm} = \mu_{\pm} \left| z_{i_{\pm}} \right| F \tag{23.17}$$

and for a 1:1 electrolyte, where $z_+ = |z_-| = z_i$:

$$\Lambda = z_i \left(\mu_+ + \mu_-\right) F \tag{23.18}$$

In the case of the conduction of ions, the product of the mobility and the electric field will give the drift velocity:

$$v = \mu E \tag{23.19}$$

Since the equivalent conductivity of an ion must be related to the ease of moving it in an electric field, that is, to the mobility, it would be predicted on the basis of Eqs. (23.15) through (23.19) that the equivalent conductivity of an electrolyte should increase as the viscosity decreases. Since the viscosity of aqueous solutions decreases as the temperature rises, it would be reasonable to expect the equivalent

Table 23.4 Comparison between the crystallographic radii and the λ^0 for selected ions

Ion	Crystallographic radius (pm)	Hydration number	$\lambda^\circ_\pm$ $(\Omega^{-1}\ \text{m}^2 \times 10^{-4})$
H_3O^+	133.0	3	349.82
Li^+	60.0	5	38.69
Na^+	95.0	4	50.11
K^+	133.0	3	73.52
Cl^-	181.0	2	76.34
Br^-	195.0	2	78.40
I^-	216.0	1	76.80

conductivity to increase with increasing temperature. This is true as Fig. 23.5 shows. Further examination of Eq. (23.15) suggests that the equivalent conductivity will fall as the ionic radius increases. This raises the same question that we previously discussed when developing a model of ion–solvent interactions: Which radius do we use, the crystallographic or solvated radius of an ion? Table 23.4 lists the crystallographic radii and equivalent ionic conductivities at infinite dilution for a variety of ions. Clearly, the conductivity falls with decreasing crystallographic radius. When the association of the hydration sheath with the ion is included, however, it appears that as the hydration sheath becomes larger the conductivity falls, as would be expected. The association of a larger hydrodynamic radius with smaller ions is expected, because, as we learned in Chapter 15, the formation of the hydration sheath is essentially electrostatic in nature. Electrostatic theory predicts a stronger electric field to be associated with the ions of smaller radii and consequently a larger association of solvent molecules and a larger hydration sheath. The behavior of ions in an electric field is consistent with this picture. Thus the mobility of the ion will be dependent on its relationship with the solvent in terms of the solvent's properties, such as viscosity, and on its association with the solvent, as in the hydration sheath.

If the considerations of ion–solvent interactions are important, what about the effects of ion–ion interactions? These interactions will be our next exploration. Before moving ahead, we note that the equivalent conductivity of the proton is exceptionally high, especially when the fact that the proton exists as a hydrated hydronium ion in solution is considered (its hydration number of three suggests that it is part of a tetrahedral array with other water molecules). We must account for this anomalously high conductivity before our treatment of ionic conductivity is complete (cf. Section 23.5).

23.5 Interionic Forces Affect Conductivity

The consideration of ion–solvent interactions provided a reasonable explanation for the differences found in the equivalent conductivities of ions at infinite dilution. Thus a parallel to our earlier discussion of the chemical potential and activity coefficient is established. Following this track, it is likely that interactions between the ions will lead to effects on the measured conductivity. We have argued that ions in solution are surrounded by a cloud of ionic countercharge (cf. Chapter 16). It is not hard to imagine that as the concentration of electrolyte in solution increases, the ionic clouds may interact with one another; that is, the migration of the ions is no longer independent of other ions. As Fig. 23.10 shows, because the net flux, and hence the conductivity, of an electrolyte solution depends on the sum of the movement of positive charge and negative ions, each moving in opposite directions, the resistance to current flux will increase as the ion clouds drag past one another. This effect of an increased viscous force due to the interaction of the ionic clouds is termed the *electrophoretic effect*. Figure 23.11 illustrates this effect.

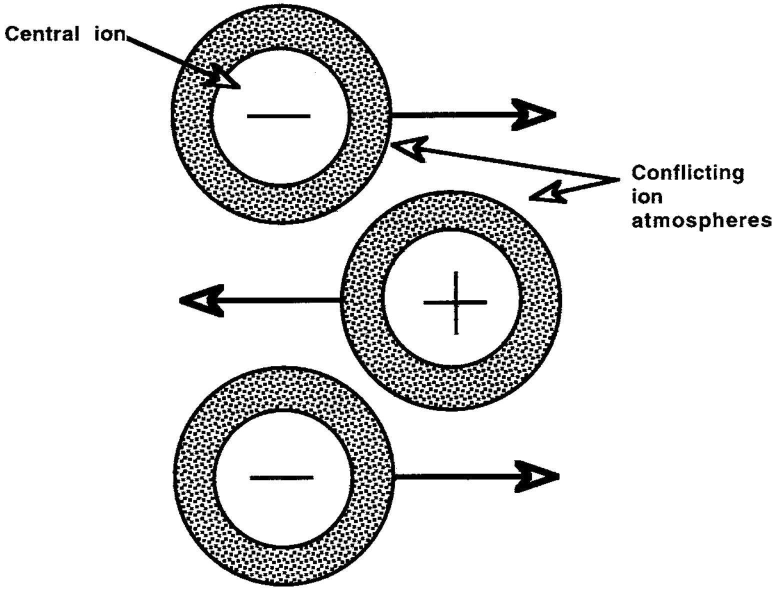


Fig. 23.11 Ionic cloud interference leads to an increased viscous force in the electrophoretic effect

The picture of the spherically symmetrical ionic cloud that was detailed earlier was developed under static conditions. What are the dynamic events when a central ion begins to move due to the electrical potential field? As the ion moves down the gradient of force, the ionic cloud also needs to move along with the ion. However, the "ionic cloud" is composed of a time-average arrangement of other ions that are drawn to the central ion. As the central ion moves, two asymmetric fronts develop. The asymmetry of the dynamic ionic cloud is a result of the finite time it takes for the old cloud to decay and the new cloud to form. Because the ions can neither dissipate behind the moving ion nor form a new cloud in front of the ion instantly, the ion experiences a separation of charge within its own cloud. The center of charge is in fact behind the ion, thus generating a local electric field that acts to retard the

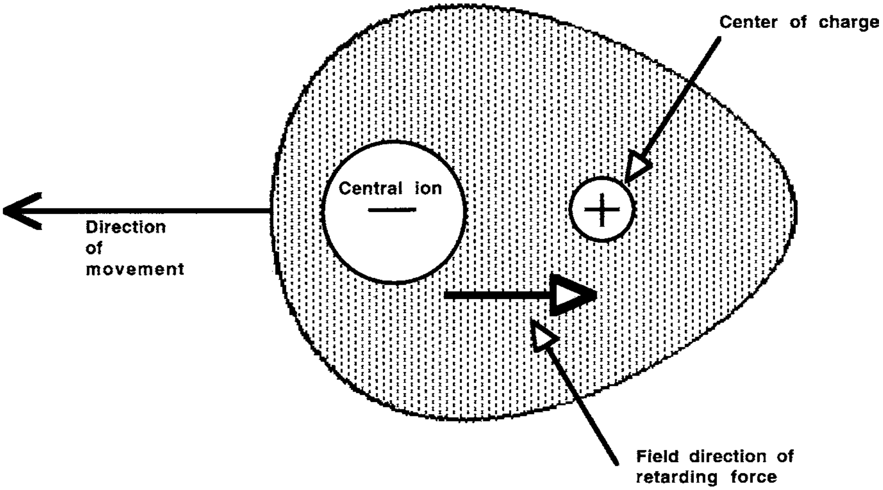


Fig. 23.12 The dynamics of the relaxation effect

movement of the ion with respect to the external field. This retarding force due to the finite time it takes for ion cloud rearrangement to occur is called the *relaxation effect*. The relaxation time for dilute solutions is on the order of 10^{-6} s. The relaxation effect is illustrated in Fig. 23.12.

Both the electrophoretic and the relaxation effects depend on the density of the ionic atmosphere and can be shown to increase with $\sqrt{c}$. This is consistent with the observations made by Kohlrausch as described earlier (cf. Section 23.3). A model theory based on these effects has been developed and is called the *Debye–Hückel–Onsager* theory. A rigorous development of the theory is beyond the scope of this volume, but the conclusion can be given. This model has the form:

$$\Lambda = \Lambda^{\mathrm{o}} - \left(A + B\Lambda^{\mathrm{o}}\right) c^{1/2} \tag{23.20}$$

Here, the constant A describes the electrophoretic effect and has the form:

$$A = \frac{z_i e_0 F}{3\pi \eta} \left(\frac{2z_i^2 e_0^2 N_A}{\varepsilon_0 \varepsilon kT} \right)^{1/2} \tag{23.21}$$

B describes the relaxation effect and can be written as

$$B = \frac{e_0^2 z_i^2 q}{24\pi \varepsilon_0 \varepsilon kT} \left(\frac{2e_0^2 z_i^2 N_A}{\pi \varepsilon_0 \varepsilon kT} \right)^{1/2} \tag{23.22}$$

where q is approximately 0.5 for a symmetrical electrolyte. This equation provides reasonable agreement with the empirical conductivity data, thus providing good evidence for the ion–ion interactions described so far. It is accurate for concentrations up to only approximately 0.002 *M* in the case of aqueous solutions of 1:1 electrolytes. Further modifications have been suggested for more concentrated solutions, but the assumptions and development of these theories are both complex and controversial and will not be discussed.

Two further lines of evidence that support the model of ion–ion interaction that we just described are worth mentioning. The first is the *Debye–Falkenhagen effect*, in which conductivities are studied with a high-frequency electric field. In a high-frequency field, the ionic cloud will not have time to form and hence the displacement of the center of charge will not occur. Without the separation of the charge, the retarding force responsible for the relaxation effect would not be expected to occur and conductivities should be higher under these conditions. Experiment confirms this effect. While the Debye–Falkenhagen effect is probably of no significant consequence in biological systems, a second line of evidence may well have biological relevance. For electric fields of low strength, i.e., 10^2–10^3 V/m, the molar conductivity of a true electrolyte is independent of field strength. However, for fields of high strength, i.e., $10^6 - 10^7$ V/m, the conductivity is found to be increased. This is the first *Wien effect.* At high field strengths the movement of an ion is so rapid that it moves an order of magnitude faster than the relaxation time of the ionic cloud formation. Consequently, the ion travels without its ionic cloud and the retarding force due to the charge separation is eliminated. It is interesting to note that the transmembrane field strengths of many cells are on the order of 10^6 V/m or greater and the ion fluxes across these membranes may well be subject to this Wien effect.

23.6 Proton Conduction Is a Special Case that Has a Mixed Mechanism

The equivalent conductivity of protons in aqueous solutions is significantly higher than would be predicted on the basis of the previous discussion. Protons have been shown to exist as hydronium ions and not as free protons in aqueous solution. As Table 23.4 shows, the ionic radius of the H_3O^+ ion is close to the radius of a potassium ion. Furthermore, the hydronium ion has been shown to exist in a tetrahedral complex as $(H_9O_4)^+$. Considering these facts our present discussion does not seem to explain the surprising mobility of the proton. Could the structure of the aqueous solvent be responsible for the high mobility? Experimental support for this idea is found by demonstrating that the equivalent conductivity of H^+ falls to the level of K^+ when the water is substituted with another solvent. What is the unique interaction between protons and water that makes this phenomenon so dramatic?

We have characterized the protons in water as part of a dynamic network of hydrogen-bonded water clusters. The existence of this extensive hydrogen-bonded structure is part of the reason that proton mobilities are so high in water. Rather than actually being physically transported down the electrical potential gradient,

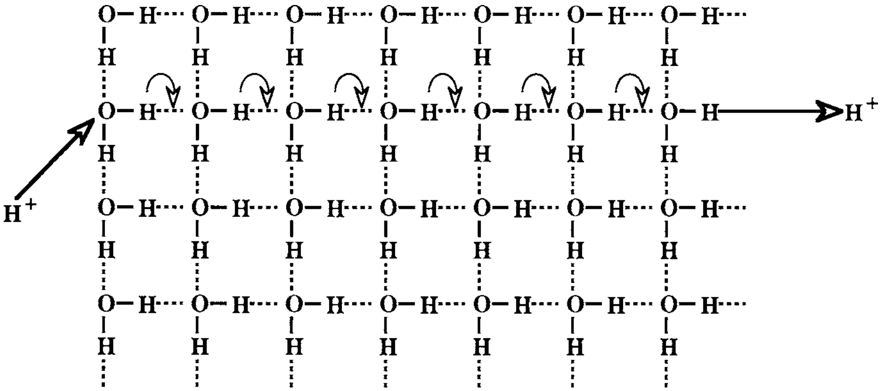


Fig. 23.13 Schematic of the movement of a proton through liquid water by a bucket-brigade mechanism

the proton is passed through the lattice from molecule to molecule with the net effect being the movement of charge. This bucket-brigade mechanism is illustrated in Fig. 23.13 and was suggested by I. Klotz. However, while liquid water and ice are both highly structured due to the hydrogen bonding of the molecules, this structure is not static. The idea that the net transport of protons occurs by a "simple" brigade mechanism must be further scrutinized.

In Section 30.2, a picture of mobility based on the ideas of mean free path and rate processes was described. It was noted that the mean free path for a particle making a jump in solution depends greatly on the structure of the solvent and that the rate or frequency of the jumps depends on the size of the energy barrier between the pre- and the post-jump coordinates. Can the exceptional mobility of the proton moving in a brigade fashion be understood in these terms? First, consider the question of the potential energy barrier that will exist between the pre- and the post-jump water molecules on the brigade line. The process of proton exchange in the brigade can be written as follows:

$$H_2O - \text{proton} \rightarrow H_2O - \text{proton}^* \rightarrow \text{proton} - H_2O \qquad (23.23)$$

This process is symmetric and it would be expected that a potential energy diagram could be drawn for it. This diagram (Fig. 23.14) would be comprised of two mirror image plots of the potential energy associated with removing a proton from an H_3O^+ molecule. The potential energy barrier associated with proton transfer is on the order of 40 kJ/mol. When this value is used in a kinetic analysis like that outlined in Chapter 30, there is reasonable agreement between the predicted and the measured mobilities. However, the analysis of proton mobility is not really complete at this point. Although the concurrence between theory and experiment is good for the proton mobility, other experimental tests suggest that the model as proposed needs further adjustment. One of these conflicts is in the prediction that the mobility of H^+ in H_2O will be significantly increased over that of D^+ in D_2O; this, it turns out,

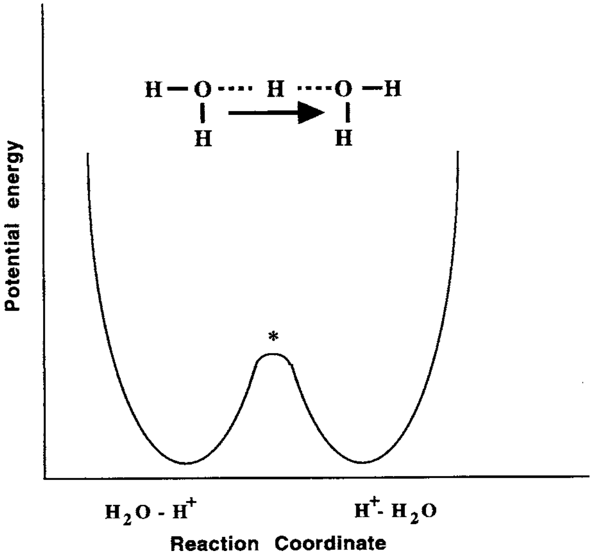


Fig. 23.14 Potential energy graph showing the potential energy barrier that occurs as a proton is abstracted from an oxygen center (*left* well) and is moved to a new acceptor center (*right* well)

is not the case (the ratio of mobilities is 1.4). A second conflict lies in the prediction that with increasing temperature the distance between the donor and the acceptor H_2O molecules will increase, leading to a higher energy barrier and hence a higher heat of activation. In fact, the energy of activation falls with increasing temperature. What further considerations should be included in this analysis?

Protons, although not as small from the standpoint of mass as electrons, are sufficiently light that they are better treated by quantum mechanics than by classical techniques. We know that one of the results of a quantum mechanical treatment of a particle in a potential energy well with finite barriers is that a distinct probability exists that the particle may be found outside the potential barrier without having climbed up and out of the well. In effect, the particle tunnels through the barrier. This leads to the physical consequence that particles may cross a potential energy barrier, even if they do not have sufficient energy to go over the barrier in classical fashion. The consideration that protons may tunnel in aqueous solution is an important one. It has been shown that tunneling is the most likely mode of transport in water and it is preferred over the classical approach given above. However, comparing the mobilities derived from the tunneling calculations with the experimental measurements leads again to trouble – the mobility predicted by a tunneling approach is simply too fast. Considerations of speed, however, are only meaningful when there is somewhere to go. If a proton can rapidly tunnel through a barrier, where does it end up? Attention must be turned to the ability of the acceptor H_2O molecule to receive a proton.

A water molecule is able to assume only a limited number of positions where its unbonded electrons can form a bond with the transferring proton. There is a finite time that is required for the water molecule to assume a position where it can receive a proton. Furthermore, since the concern is with conduction of the proton, the same water molecule must now reorient and assume a position in which it can pass on the proton to another acceptor molecule. Even if the actual transfer is accomplished by a rapid tunneling process, the rate of transfer will depend on the time necessary to position and reposition the acceptor/donor water molecules along the way. If calculations are made based on the random movement of the H_2O molecules induced by thermal forces, the reorientation occurs too slowly to account for the experimentally found mobility. However, it is not reasonable to treat this system as being dependent only on thermal forces for its interactions. An H_3O^+ molecule is positively charged and will exert an attractive coulombic force on the partial negative charge of the acceptor dipole molecule. This coulombic interaction will act to increase the speed of reorientation; hence the generation of a local electric field will assist the reorientation of the acceptor molecule. This orienting force significantly increases the rate of reorientation over that found for a strictly thermally induced motion and allows for a mobility consistent with that found experimentally. The complete picture of proton conduction in H_2O is a combination of fast tunneling through, rather than over, potential energy barriers but with a rate-limiting step, the orientation of the acceptor molecule. While the correct orientation of the acceptor molecule is the rate-limiting step, this step is accelerated by the ion–dipole interaction of the approaching hydronium ion whose electrostatic field assists and speeds the reorientation of the acceptor molecule. Not only does this model give proton mobilities that are consistent with experiment, but also it predicts the surprising fall of activation energy with increased temperature and gives the correct ratio of H^+/D^+ mobility (incorrectly forecast by the earlier model). The activation energy drop is predicted, because in the reorientation model the rate of orientation will be related to the number of hydrogen bonds that must be broken and re-formed. As the temperature increases, the number of hydrogen bonds is decreased, reorientation is easier, and the activation energy falls.

Further Reading

General

Bockris J.O'M. and Reddy A.K.N. (1998) *Modern Electrochemistry*, 2nd edition, Volume 1. Kluwer Academic/Plenum, New York.

Edsall J.T. and Wyman J. (1958) *Biophysical Chemistry*. Academic, New York.

Harned H.S. and Owen B.B. (1950) *The Physical Chemistry of Electrolytic Solutions*. Reinhold, New York.

Waldram, J.R. (1985) *The Theory of Thermodynamics*. Cambridge University Press, Cambridge.

Aqueous Proton Conduction

Conway B.E. (1964) "Proton Solvation and Proton Transfer Processes in Solution." In: J.O'M. Bockris and B.E. Conway (eds.), *Modern Aspects of Electrochemistry*, Volume 3. Butterworths, London.

Conway B.E. and Bockris J.O'M. (1959) On the theory of proton conductance in water and ice, *J. Chem. Phys.*, **31**:1133–1134.

Conway B.E., Bockris J.O'M., and Linton H. (1956) Proton conductance and the existence of the H_3O ion, *J. Chem Phys.*, **24**:834–850.

Problem Sets

1. Determine the equivalent conductivity at infinite dilution of the following solutions using Kohlrausch's law of independent migration.
   (a) NaCl
   (b) KCl
   (c) H_2O
   (d) $CaCl_2$
   (e) $Ca(OH)_2$
2. When most people lie or make an untrue statement their sympathetic nervous system becomes active (fight or flight response). This leads to increased sweat secretion and thus decreased skin resistance. In a lie detector, a current is passed between two silver chloride electrodes held at 5 V. A current of 4.2 mA is measured when the subject gives his name. Later in the interview he lies and a current of 8.9 mA is measured. What is the skin resistance in the (a) honest and (b) dishonest state?
3. The brain consumes 20% of the oxygen carried by the blood. The heart pumps 5 l/min of blood and the oxygen content of the blood is $pO_2 = 120$ mmHg.
   (a) Using Faraday's law, how many amperes does the brain produce in an hour to reduce the oxygen?
   (b) Since the voltage drop from $NADH + O_2 \rightarrow NAD + H_2O = 1.12$ V, what is the power output of the brain per hour?
4. Calculate the drift velocity for the listed ions under the field condition of the transmembrane potentials before depolarization [–80 mV], at full discharge [+50 mV] and during hyperpolarization [–90 mV]. Assume the ion is unhydrated.
   (a) Na^+
   (b) Cl^+
   (c) K^+
   (d) $HCO_3{}^-$
   (e) H^+
5. Calculate the drift velocities for the ions in problem 4 when the ions are hydrated.

Chapter 24
Forces Across Membranes

Contents

24.1 Energetics and Force in Membranes

Much of the chemistry and functional physiology of the cell, including energy production, protein synthesis, hormone and antigen binding, stimulus–response coupling, and nutrient adsorption, occurs at the cell membrane. The membrane is an anisotropic non-homogeneous matrix of lipids, proteins, and, in some cases, carbohydrates that is in intimate contact with aqueous-dominated interphases. The nature and treatment of aqueous solutions, membrane properties, and interphase

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_24,

structure have all been presented in previous chapters. The task ahead is to highlight this knowledge by examining some limited aspects of the behavior or action of biochemical systems. In this chapter we focus on describing the forces operating at and across the membrane. Then we will examine the role of the membrane in modulating the flow of materials.

24.2 The Donnan Equilibrium Is Determined by a Balance Between Chemical and Electrical Potential in a Two-Phase System

The determination of the equilibrium in a homogeneous solution of several components and the approach to equilibrium via diffusion and conduction have been discussed. The addition of a semipermeable membrane will create a heterogeneous system in which the equilibrium between two phases must be considered (Fig. 24.1). This membrane is freely permeable to simple electrolytes such as Na^+ or Cl^- but is impermeable to a colloidal electrolyte. In our example, a protein of molecular weight 82,000 Da will be the polyelectrolyte to which the membrane is impermeable. Initially, aqueous solutions of only NaCl are added to each side. Since the membrane is freely permeable to each component of the solutions, the chemical potential of each component in each phase will be equal at equilibrium, even if the initial concentrations of these solutions are different. The following equations can be written to represent the equilibrium condition:

$$\mu^{\mathrm{A}}_{\mathrm{Na}^+} = \mu^{\mathrm{B}}_{\mathrm{Na}^+} \tag{24.1}$$

$$\mu^{\mathrm{A}}_{\mathrm{Cl}^-} = \mu^{\mathrm{B}}_{\mathrm{Cl}^-} \tag{24.2}$$

$$\mu^{\mathrm{A}}_{\mathrm{H_2O}} = \mu^{\mathrm{B}}_{\mathrm{H_2O}} \tag{24.3}$$

Assuming ideality in this example, it can also be written that the concentrations of Na^+ and Cl^- are equal in each compartment; this fulfills the requirement for electroneutrality:

$$C^{\mathrm{A}}_{\mathrm{Na}^+} = C^{\mathrm{A}}_{\mathrm{Cl}^-} = C^{\mathrm{B}}_{\mathrm{Na}^+} = C^{\mathrm{B}}_{\mathrm{Cl}^-} \tag{24.4}$$

The equilibrium is now perturbed by the addition of the polyelectrolyte to side A. Since virtually all biological macromolecules have some polyelectrolyte nature, this case is a thoroughly general one. This protein will be given a net cationic charge. Through the use of a thought experiment, the charge on this protein may be switched on or off at will. Initially, the charge is turned off. What is the effect of the addition of the uncharged protein? The addition of the protein to side A leads to a fall in the mole fractions of Na^+, Cl^-, and H_2O with a concurrent drop in their activities. The osmotic pressure on side A necessary to raise the activity of the permeable

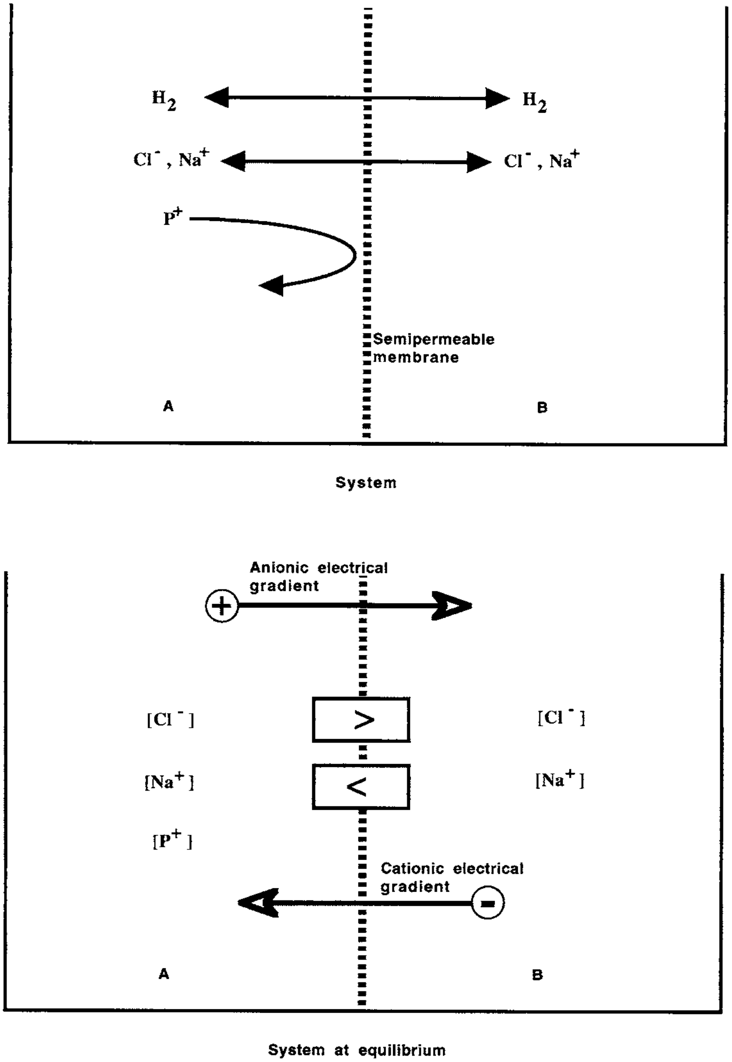


Fig. 24.1 The Gibbs–Donnan equilibrium describes the case of two phases separated by a semipermeable membrane that allows free passage of H_2O and ions, but not macromolecular polyelectrolytes

components is now higher because of the addition of the protein. If the system is allowed to proceed toward equilibrium, there will be a shift in the concentrations of the permeable components just as was described in the case of the equilibrium dialysis. Ultimately, there will be a movement of water, Na^+, and Cl^- toward side A so that the chemical potentials in each phase are once again the same.

Now, the charges on the protein are switched on. What is the effect? The phases are no longer at equilibrium because electroneutrality has been violated. Side A now contains a charge that is not able to freely partition between the two solutions. What is the effect on the final equilibrium for this system? In this case, the presence of an unbalanced cationic charge on side A makes it electrically positive with respect to side B. There is an electrical driving force caused by the presence of the polyelectrolyte, and also a chemical potential driving force to push the macromolecule down its concentration gradient. The electric field causes the repulsion of the permeable cations (Na^+) from side A and attracts the permeable anions (Cl^-) from side B. The equilibrium direction attempts to reestablish electroneutrality at the expense of establishing a chemical potential gradient. As the Na^+ and Cl^- ions accumulate in different concentrations across the membrane in response to the electric field, a chemical potential difference is generated where [in terms of Eq. (24.4)]:

$$C^{\mathrm{A}}_{\mathrm{Na}^+} < C^{\mathrm{B}}_{\mathrm{Na}^+} \tag{24.5}$$

$$C^{\mathrm{A}}_{\mathrm{Cl}^-} > C^{\mathrm{B}}_{\mathrm{Cl}^-} \tag{24.6}$$

Finally, at equilibrium, there will be offsetting electrical and chemical potentials. Although ΔG will be 0, there will be a persistent electrical gradient driving the anions to move by conduction to side A and for the cations to move by conduction to side B. On the other hand, there will be a persistent chemical potential gradient driving the anions to move by diffusion to side B and the cations to move by diffusion to side A. This electrochemical gradient is called the *Gibbs–Donnan* or *Donnan potential*.

The mathematical derivation of the Gibbs–Donnan effect is as follows. Ideality is assumed in this derivation. The salt is denoted as M^+N^-, the protein as P^+, and the charge on the protein as z_i. The requirement for chemical potential equality is given by

$$\mu^{\mathrm{A}}_{M^+} = \mu^{\mathrm{A}}_{N^-} = \mu^{\mathrm{B}}_{M^+} = \mu^{\mathrm{B}}_{N^-} \tag{24.7}$$

The equation for the electroneutrality condition on side A is given by

$$Z_{\mathrm{i}} C^{\mathrm{A}}_{P^+} + C^{\mathrm{A}}_{M^-} = C^{\mathrm{A}}_{N^-} \tag{24.8}$$

The equation for electroneutrality on side B is given by

$$C^{\mathrm{B}}_{M^+} = C^{\mathrm{B}}_{N^-} \tag{24.9}$$

Equations (24.7, 24.8, 24.9) are combined, with the following result:

$$C^{\mathrm{B}}_{M^+} = C^{\mathrm{A}}_{M^+} \left[1 + \frac{Z_{\mathrm{i}} C^{\mathrm{A}}_{P^+}}{C^{\mathrm{A}}_{M^+}} \right]^{1/2} \tag{24.10}$$

and:

$$C_{N^-}^{\mathrm{B}} = C_{N^-}^{\mathrm{A}} \left[1 - \frac{Z_{\mathrm{i}} C_{\mathrm{P}^+}^{\mathrm{A}}}{C_{N^-}^{\mathrm{A}}} \right]^{1/2} \tag{24.11}$$

This is the mathematical equivalent of the earlier thought experiment. There will be a higher concentration of the cation on side B and a higher concentration of the anion on side A at equilibrium. The free energy secondary to the chemical potential gradient will be offset exactly by an opposite electrical potential gradient. The effect of the Donnan equilibrium on the osmotic pressure in the system is often important and can be significant in cellular systems.

24.3 Electric Fields Across Membranes Are of Substantial Magnitude

Biological membranes have interesting electrical properties that can be derived from measurements of the dimensions and electrical capacitance of the membranes. Biological membranes have a capacitance of approximately 1 μF/cm^2 and are about 7.5 nm thick. The measured transmembrane potential is on the order of 0.050–0.100 V. The bilayer membrane can be considered roughly as a parallel plate capacitor. Therefore, the electric field strength across the membrane will be between 7×10^6 and 1×10^7 V/m. The cell membrane is capable therefore of sustaining a large potential without breaking down, a property quantitated by a measure called the *dielectric strength*. The dielectric constant of the membrane can also be estimated from the formula for capacitance of a parallel plate capacitor:

$$\mathrm{C} = \frac{\varepsilon \varepsilon_0 A}{d} \tag{24.12}$$

where C is the capacitance, A is the cross-sectional area, and d is the distance separating the two plates. ε is the dielectric constant and ε_0 is the permittivity of free space. Using this formula and the values given above, the membrane has a calculated dielectric constant of approximately 8.

24.3.1 Diffusion and Concentration Potentials Are Components of the Transmembrane Potential

Two NaCl solutions of different concentrations are brought into electrical contact either through a salt bridge or a membrane, and a voltmeter of very high impedance is used to measure the potential difference between two Ag/AgCl electrodes immersed in the solutions (Fig. 24.2). A potential will be found between

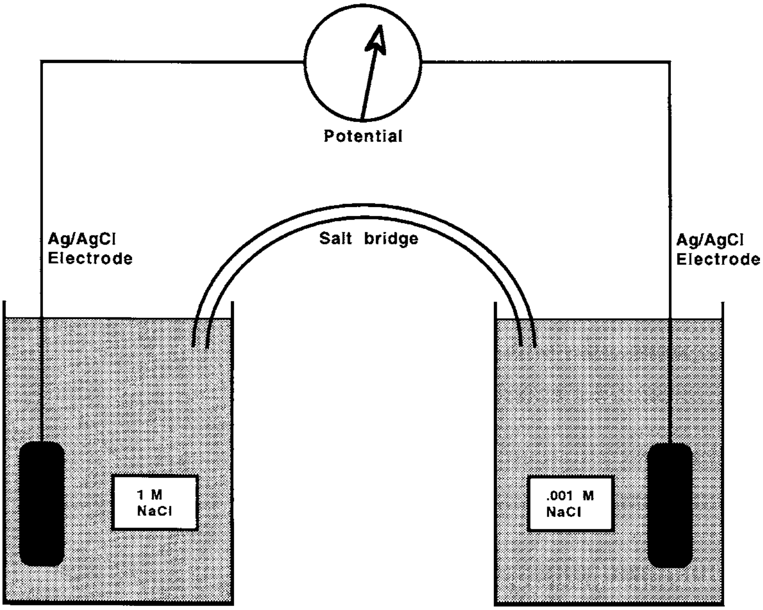


Fig. 24.2 An example of a system generating a concentration potential

the two solutions. The magnitude of the potential will be given by the Nernst equation:

$$E = -2.303\frac{RT}{nF}\log\frac{C_2}{C_1} \tag{24.13}$$

The emf is generated by the concentration difference between the two solutions and represents the difference in chemical potential between them, as sensed by the electrodes, which in fact are responding to the difference in chloride ion concentration. A potential generated in this fashion is called a *concentration potential*.

A different kind of potential can be shown in a system where a semipermeable membrane that will allow the free exchange of water and ions is placed between two NaCl solutions of different concentrations and the ions are allowed to progress down their chemical potential gradients toward equilibrium. Each ion will move down its gradient toward equilibrium but, because the mobilities of Na^+ and Cl^- are different, the Na^+ ions will diffuse faster than the Cl^- ions. As a consequence, there will be a front of positive charge that moves ahead of the diffusing negative charge. There will be a charge separation across the membrane with the more concentrated solution negative with respect to the more dilute solution. A potential of this type, illustrated in Fig. 24.3, is called a *diffusion potential*.

The diffusion potential depends on both the concentration difference between the two phases and the difference in mobilities of the ions carrying the charge:

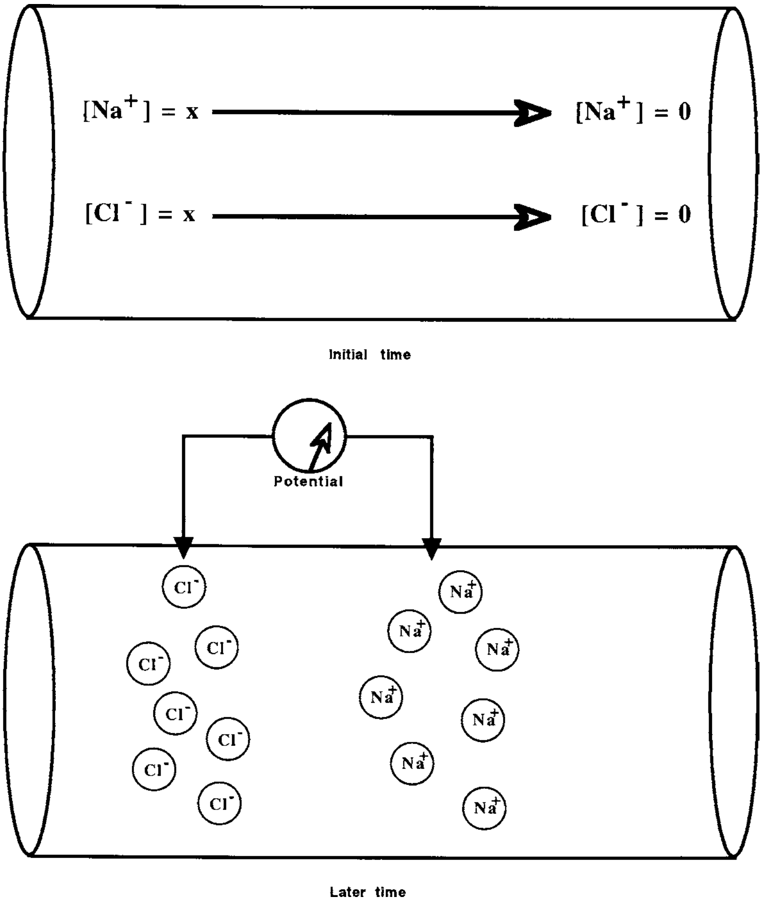


Fig. 24.3 The diffusion potential results from the different mobilities of the ions as they move down a chemical potential gradient

$$E = -\left[\frac{\mu_{+} - \mu_{-}}{\mu_{+} + \mu_{-}}\right] 2.303 \frac{RT}{nF} \log \frac{C_2}{C_1} \tag{24.14}$$

When the mobilities of the anions and cations are identical, there will be no diffusion potential. The emf generated by a diffusion potential reaches a steady state because as the charge separation increases, the potential field that is generated between the positive and negative fronts causes the leading positive ions to be retarded and the lagging negative ions to be accelerated. The concept of the diffusion potential is important because it helps to explain the physical basis for the generation of the transmembrane potentials in biological systems.

24.3.2 The Goldman Constant Field Equation Is an Expression Useful for Quantitative Description of the Biological Electrochemical Potential

In biological systems, electrical potentials are generated across membranes because of a combination of chemical potential gradients that are established by active ion transport systems that are coupled with a differential permeability of the membrane to the movement or mobility of the ions down their chemical potential gradients. Although the molecular mechanism is not the same as described for the diffusion potential, this is a similar phenomenon and can be treated thermodynamically in much the same fashion. The Goldman equation is based on the principles of *electrodiffusion*. Other treatments, notably ones based on rate theory, have also been proposed. The rate theory approach will not be described here but may be found discussed in sources listed in the further reading.

Equation (24.14), which describes electrodiffusion and is applicable for any freely permeable ion which distributes across the membrane according to the Gibbs–Donnan equilibrium, is the starting point for this analysis. The equation has been generalized for mammalian cells, which, in general, are much less permeable to cations than to anions. The principal membrane potential contributions come from the Na^+ and K^+ gradients, which are maintained by active pumps, while Cl^- and $\mathrm{HCO_3}^-$ anions have much larger membrane permeabilities and distribute according to the Gibbs–Donnan equilibrium. Generally, this is written as

$$\frac{\left[\mathrm{Cl}^-_{\mathrm{in}}\right]}{\left[\mathrm{Cl}^-_{\mathrm{out}}\right]}=\frac{\left[\mathrm{OH}^-_{\mathrm{in}}\right]}{\left[\mathrm{OH}^-_{\mathrm{out}}\right]}=\frac{\left[\mathrm{HCO}^-_{3\mathrm{in}}\right]}{\left[\mathrm{HCO}^-_{3\mathrm{out}}\right]} \tag{24.15}$$

Neither H^+ nor Ca^{2+} ions contribute measurably to the plasma membrane potentials of such cells, even though membrane transport systems and pumps driven in part by these cations exist. The generalized equation, known as the *Goldman–Hodgkin–Katz constant field equation*, expresses the transmembrane potential at equilibrium in terms of the specific membrane permeabilities for each ion, P_i, and their intra- and extracellular concentrations, $\mathrm{C}_{\mathrm{in}}^{M^+}$ and $\mathrm{C}_{\mathrm{out}}^{M^+}$, respectively, for cations and $\mathrm{C}_{\mathrm{in}}^{N^+}$ and $\mathrm{C}_{\mathrm{out}}^{N^-}$ respectively for anions, and the change z_i on the macromolecular species of concentration, C_p. This treatment incorporates several assumptions including the following:

1) the partial permeability of the membrane to charged species;
2) the uniformity of charge distribution across the membrane;
3) the net equality of charge flux across the membrane;
4) the applicability of the Donnan equilibrium for each ionic species to which the membrane is permeable enough to permit free distribution across the membrane; and
5) the absence of an electrogenic pump and its resultant driving force.

The transmembrane potential, $\Delta\Psi$, can then be written as

$$\Delta\Psi = \frac{RT}{F} 1\text{n} \frac{\sum C_{\text{in}}^{N-}}{\sum C_{\text{out}}^{N-}} = \frac{RT}{F} 1\text{n} \frac{\sum C_{\text{out}}^{M+} + Z_{\text{i}} C_P}{\sum C_{\text{in}}^{M+} + Z_{\text{i}} C_P} \qquad (24.16)$$

The limited membrane permeability P_i for any species i is defined as

$$P_i = \frac{RT}{d} \mu_i b_i \qquad (24.17)$$

where d is the thickness of the membrane, μ_i is the mobility of the ion in question, and b_i is a measure of the ease of distribution of the ion between the outside and the inside of the cell. The assumption is that as the thickness of the membrane varies, the permeability varies in a constant fashion. The Goldman–Hodgkin–Katz constant field equation is the result of

$$\Delta\Psi = \frac{RT}{F} 1\text{n} \frac{\sum P_i C_{\text{out}}^{M+} + \sum P_i C_{\text{in}}^{N-}}{\sum P_i C_{\text{in}}^{M+} + \sum P_i C_{\text{out}}^{N-}} \qquad (24.18)$$

While this equation holds for most mammalian cells at or near physiologic conditions, it tends to fail when such cells are suspended in very non-physiologic buffers (e.g., external $K^+ >> 50$ mM). The Goldman equation has a number of significant limitations, as well as useful extensions including to ions of other valences and electrogenic pumps. Discussion of these modifications can be found in references listed at the end of the chapter.

24.4 Electrostatic Profiles of the Membrane Are Potential Energy Surfaces Describing Forces in the Vicinity of Membranes

If we consider the overall electrostatic potential profile of the bilayer lipid membrane, it is clear that several sources of charge generate the field operating around and through the membrane. One of it is the *transmembrane potential* that is generated from the unequal partitioning of ionic charge. A second is the *surface charge* of the membrane itself derived from the ionizable moieties of the components of the membrane. Each of these components is usually treated as its generative charge is found at the surface. An examination of the structure of the electrified interphase in the direction of the aqueous solution was developed in Chapter 20. The potential field across the membrane itself is important in biological systems because of its influence on the biologically active components both resident and transient to the membrane.

The metal–electrolyte interphase described in Chapter 20 suffers from the fact that charge in metal conductors is carried in the quantum mechanically delocalized conduction bands of the metal. This leads to some rather specific electronic interactions with the electrolyte that is not mechanistically similar to the case of all of the charge being provided by electrolytes. A simple model of the interface, that is sometimes considered instead of the metal–electrolyte systems described in Chapter 20, is one that contains only ions as the charge carriers at the interface. This model is called the *interface of two immiscible electrolyte solutions* (*ITIES*). In this model, the electrolytes partition across the interface depending on a relative difference in their hydrophilicity. A charge separation occurs and in such a model, the arrangement of the ions in the interphases on both sides of the interface is given as a Gouy–Chapman type diffuse layer. The surface charge density for a phase is related to surface potential ψ_{s} as given in the last chapter:

$$\sigma_{\mathrm{W}} = (8\varepsilon_{\mathrm{o}}\varepsilon_{\mathrm{W}}RTc_x)^{1/2}\sinh\frac{zF\Psi_{\mathrm{S-W}}}{2RT} \tag{24.19}$$

Here we write the expression for the aqueous phase with ε_{W} representing the permittivity of water and c_x the concentration of the z:z valent electrolyte. Applying the exact equation for the solution of $\psi_{\mathrm{s\text{-}w}}$ at a surface we can write as

$$\Psi_{\mathrm{S-W}} = \frac{RT}{zF}1\mathrm{n}\left[\frac{\sigma_{\mathrm{W}}}{(8\varepsilon_{\mathrm{o}}\varepsilon_{\mathrm{W}}RTc_w)^{1/2}} + \frac{\sigma_{\mathrm{W}}^2}{(8\varepsilon_{\mathrm{o}}\varepsilon_{\mathrm{W}}RTc_w + 1)^{1/2}}\right] \tag{24.20}$$

The overall electrical double layer is electroneutral. The electrical potential difference between the two phases (aqueous–organic = w-o) will be:

$$\phi_{\mathrm{w-o}} = \psi_{\mathrm{s-w}} - \psi_{\mathrm{s-o}} \tag{24.21}$$

Next consider the interface at an aqueous and liquid hydrocarbon interface. The liquid hydrocarbon is a low-permittivity solvent and when any hydrophobic ion penetrates into it, ion pairing will instantly occur. Thus the only charge separation will be generated by the adsorption of amphiphilic ions at the interface. These interactions were described earlier as the actions of surfactants. The coverage of the interface surface by these surfactant molecules is limited. The fraction of the surface covered (Γ_i, in mol/m^2) compared to the maximum possible coverage (Γ_{i-m}) is given by

$$\Theta = \frac{\Gamma_i}{\Gamma_{i-m}} \tag{24.22}$$

At constant temperature and independent site occupation, the coverage is given by the Langmuir isotherm:

$$\Gamma_i = \frac{\Gamma_{i-m}\beta_i c_i}{(1 + \beta_i c_i)} \tag{24.23}$$

where β_i is the absorption coefficient, similar to the equilibrium constant, and c_i is the concentration of the absorbant in the bulk solvent.

However, since the absorbed particles will electrostatically interact with previously absorbed charged particles, the Frumkin isotherm can be used to account for nearest neighbor interactions:

$$\beta_i c_i = \frac{\Gamma_i}{\Gamma_{i-m} - \Gamma_i} \mathrm{e}^{(-2A\Gamma_i)} \tag{24.24}$$

A is an interaction coefficient with the following form where φ is the interaction energy between one molecular pair and N_A is the number of nearest neighbors interacting with a single molecule on a completely covered surface:

$$A = \frac{N_a \varphi}{2kT}$$

A will be greater than 0 for attractive interactions and A less than 0 for repulsive interactions. When A equals 0 there is no nearest neighbor interaction and the Frumkin isotherm will reduce to the Langmuir expression. The surface charge, σ, is directly related to Γ_i if the only absorbing species is the ionic surfactant with charge z_i:

$$\sigma = \Gamma_i z_i F \tag{24.25}$$

The surface potential can then be found by Eq. (24.20).

This model has certain characteristics of the bilayer membrane but biological membranes contain the dipoles that result from the charged head groups of the phospholipids as well as the interaction of the aqueous ions forming ion pairs with the polar head groups. The surface potential therefore has components derived from the ionic charge and the dipoles:

$$\Psi_{\mathrm{S}} = \Psi_{\text{S-ion}} + \Psi_{\text{dipole}} \tag{24.26}$$

We know that the potential derived from dipoles is strongly dependent on the geometry of the dipole in space and the physical organization of the polar head groups will variably influence ψ_{S}. It is also worth reconsidering the assumptions underlying the Langmuir and Frumkin adsorption isotherms in the context of ion adsorption to a complex biological membrane. The Langmuir isotherm considers the fundamental adsorption step but no interaction energies: This is an unrealistic abstraction. The Frumkin isotherm considers nearest neighbor interactions and can represent the electrical interactions associated with ion adsorption, but the nearest neighbor treatment does not fully capture the long-range interaction inherent in ion–ion interactions and leads to substantial problems. Other adsorption isotherms exist such as the Temkin isotherm that takes into account the idea of heterogeneity or different energies of adsorption interaction across the surface (but neglects the nearest-neighbor lateral interactions); the Conway–Angerstein–Kozlowka isotherm that includes terms to account for the partial transfer of charge between the surface and the adsorbed molecule; the Habib–Bockris isotherm that accounts for displacement of solvent molecules (a Flory–Huggins treatment) and includes lateral interactions between the adsorbed species of both short and long range. In terms of ionic adsorption, the Habib–Bockris isotherm is probably the most successful of these formulations. Details of these isotherms can be found in the references listed at the end of this chapter.

Given the adsorption effects of ions, organic molecules, and aqueous dipoles, the ITIES model does not fully describe the interphase near a cell membrane. Furthermore, it does not account for the active generation of the membrane potential through the action of ion pumps known to act at the membrane. The surface charge has a magnitude of 10^{-1} to 10^{-3} C/m and is of mixed cationic and anionic charge groups. The net charge is anionic, with approximately 10^{7} to 10^{8} anionic charges per cell and 10^{6} to 10^{7} cationic charges per cell. Neither a simple nor a complete model yet exists of the electrified interphase of cellular systems. A number of possible potential profiles across membranes are illustrated in Fig. 24.4.

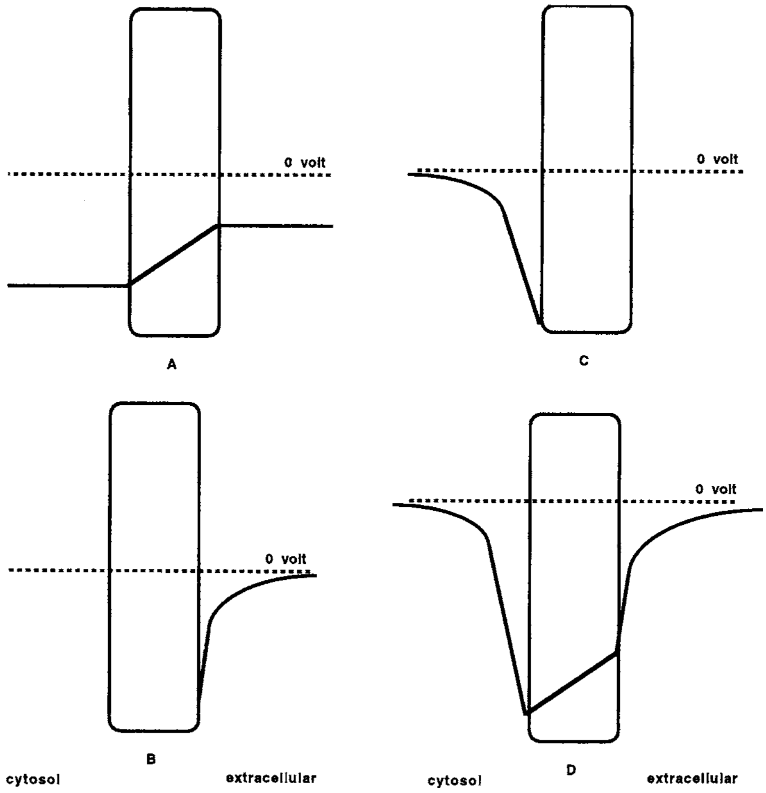


Fig. 24.4 Model electrostatic profiles across membranes. In panel (**a**), the electrostatic profile due strictly to the transmembrane potential is shown, and the membrane is treated as a dielectric continuum with $e = 8$; in panel (**b**), the electrostatic profile due to the extracellular surface of the plasma membrane is drawn, based on the surface charge assumptions given in the text; in panel (**c**), the profile due to the cytoplasmic side is shown; finally, in panel (**d**), a total picture of the plasma membrane electrostatic picture is drawn based on the principle of superposition of each of the previous curves

Ultimately the model of the biological membrane as non-homogeneous is useful because the description of the behavior of the cell membrane in terms of its macroscopic properties loses much of the detailed mechanistic information that concerns the biological worker. Like the methods of the self-consistent field in computational chemistry the use of the macroscopic description can provide the background potentials in which more detailed microscopic fields can be calculated and successive approximations will eventually lead to a much more complete picture of the biological system.

24.5 The Electrochemical Potential Is a Thermodynamic Treatment of the Gradients Across a Cellular Membrane

It is apparent that the movement of molecules into and out of the cell or across a cellular membrane will depend on a combination of gradients. This book has focused on the chemical and electrical potentials, which are so tightly coupled to one another that the combined free energy of their gradients is called the *electrochemical potential*. For the *i*th ion in the aqueous phase, the electrochemical potential (μ_i) is

$$\mu_i = \mu_i^{\mathrm{o}} + RT \ln a_i + z_i FE \tag{24.27}$$

Thus the overall electrochemical field a component of a system experiences is a combination of the vector fields of chemical and electrical potential. At equilibrium, the net force on the component will be zero but a given component may well experience significant electrical and chemical field vectors at that point in state space. When these forces are unbalanced, net movement of the component is the result. This the source of the electrochemical potential gradient that causes transport in many important biological systems.

24.6 Transport Through the Lipid Bilayer of Different Molecules Requires Various Mechanisms

24.6.1 Modes of Transport Include Passive, Facilitated, and Active Processes

Transport across a membrane is usually classified as either passive, facilitated, or active. Passive transport is the diffusion of a component down a chemical potential gradient. We explored the forces and kinetics of diffusion in Chapter 26. In terms of a diffusion through a membrane the rate of flux through the membrane will be related to the movement of the component from the bulk solution onto the surface of the membrane. Then the movement through the membrane will depend on the diffusion coefficient for that component in the membrane milieu. Finally, there must be a desorption of the component from the membrane into the other bulk phase that was separated by the membrane. Each of these conditions has been described in our treatment to this point.

Facilitated transport is needed for some molecules whose solubility in the lipid phase is very low. Hence transport of these molecules will be very slow. Charged species such as ions exhibit this behavior. Their movement across the lipid barrier can be facilitated if they are carried by an amphipathic molecule acting much the same as the detergents we have already discussed. Removal of the water molecules from the hydration sheaths around inorganic ions is a process requiring a significant amount of energy. This accounts for the co-transport of H_2O across biological membranes with a number of charged species. It also accounts for the specificity of certain compounds which act as transporters of ions across the membrane, such as the ionophore valinomycin. Note that this usage of the term ionophore is distinct from the electrochemical definition. In biological studies, the term ionophore is most commonly used to refer to compounds such as valinomycin, ionomycin, and nigericin which act to allow transport of certain ionic species across membranes. As shown in Fig. 24.5, valinomycin is a cyclic depsipeptide which folds so that all of its carbonyl groups point toward the center of a sphere. Its shape is like the seam on a tennis ball, having the depsipeptide as a backbone and the nonpolar side chains pointing toward the surface of the sphere. Its exterior is hydrophobic and lipophilic, and its interior is quite hydrophilic. Lipid membranes are freely permeable to valinomycin, whether its interior space contains a molecule of H_2O or an ion. The space within valinomycin is too small to accommodate any hydrated ion, and the compound can therefore act as a transmembrane transporter only if the energy required to strip the solvating water from the ion is compensated by the energy of binding to the carrier. Only for K^+ is such an interaction energetically favorable.

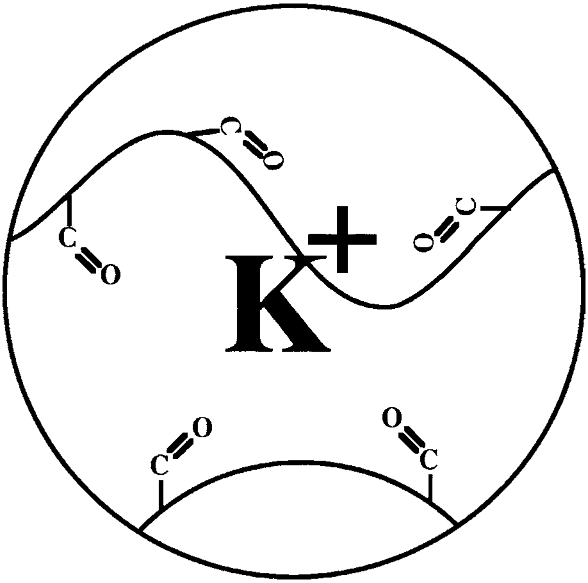


Fig. 24.5 Cartoon of the functional structure of valinomycin. The peptidyl backbone (L-lactic acid-L-valine-D-hydroxyisovalerate-D-valine) is cyclic. The carbonyls all point toward the center of the spheroid in which the potassium ion is caged. The less polar surface makes the complex soluble in the lipid membrane

Smaller ions have a greater energy of solvation and a larger separation distance and therefore have a weaker interaction with the oxygen atom of the carbonyl groups. Ions larger than K^+ cannot be accommodated in the limited intraspherical space. Therefore, valinomycin is an ionophore which is virtually specific for K^+ transport across membranes and is indeed the only known ionophore which exhibits such complete selectivity toward one ionic species.

In certain cases the net flow of a species will be against its electrochemical gradient. This type of process requires that energy be expended at the membrane to do work against the free energy gradient. This process is active transport. In biological systems such molecular pumps are found for a variety of ions such as the Na^+–K^+ pump, which derives its energy for the work cycle from the hydrolysis of ATP. The thermodynamic treatment of such a coupled system in a mixed heterogeneous electrochemical field has been the subject of the last three sections.

24.6.2 Water Transport Through a Lipid Phase Involves Passive and Pore Specific Mechanisms

The passage of any polar entity across a lipid bilayer is severely restricted and requires either the presence of pores or channels in the membrane or the existence of enough kinetic energy to drive the molecule through random (and transiently present) spaces between lipid molecules. In a mammalian membrane, transmembrane proteins can act as pores or as passive channels through which water, alone or in concert with other small molecules, can be brought across the membrane. The gradient for such passive transport will be dependent on the relative activity of H_2O in the extracellular milieu versus that in the cytoplasm. Under physiologic conditions, the driving force will clearly be inward, and, were it not for the osmotic pressure and the existence of counteracting forces attributable to cation gradients, the cell would swell and eventually burst. Therefore, in the overall free energy equation describing the movement of molecules through cell membranes, the terms involving pressure, chemical potential, surface area, and charge movement must all be considered.

Changes in cellular water content are a significant factor in the biological and clinical behavior of cells. Shrinkage of a cell due to net H_2O efflux can change the activity coefficient of the residual H_2O or change the degree of solvation of intracellular components, both cytoplasmic and membrane bound. In comparison to its permeability to other small solutes such as urea, the cell membrane as a whole is quite permeable to water. In the erythrocyte, the diffusion rate of water through the membrane is a factor of 10^2 times higher than that of urea, even though the membrane is highly permeable to urea. The exchange rate of water between the inside and outside of an erythrocyte is quite rapid, on the order of fractions of a second. In spite of the relatively high permeability of the membrane to water, a significant barrier does in fact exist. In the absence of a membrane, the relative diffusion rate of water is higher by a factor of 10^4 or 10^5.

Water can enter a cell through several mechanisms: passive diffusion directly through the lipid bilayer, passive diffusion through transmembrane protein channels or pores, or co-transport with other entities (ions or uncharged polar species). All three modes of permeation are gradient driven. Although water is co-transported across a membrane with many ions or neutral molecules via channels or carriers, it can be carried alone through any of these modes of transport.

Model systems for the passive permeation of water into intracellular spaces are easily developed. They take account of the fact that, of the four possible modes of membrane traversal by water (osmosis, pinocytosis, electro-osmosis, and vacuole contraction), osmosis is the predominant one. Implicit in all models for osmotically driven water permeation are the following assumptions:

1) The water flux is controlled at the membrane.
2) The membrane is semipermeable, passing water but no solutes.
3) The cell is at equilibrium with its surroundings.
4) The protein components of the cell interior (to which the membrane is impermeable) behave as if they were in a cell-free aqueous milieu.

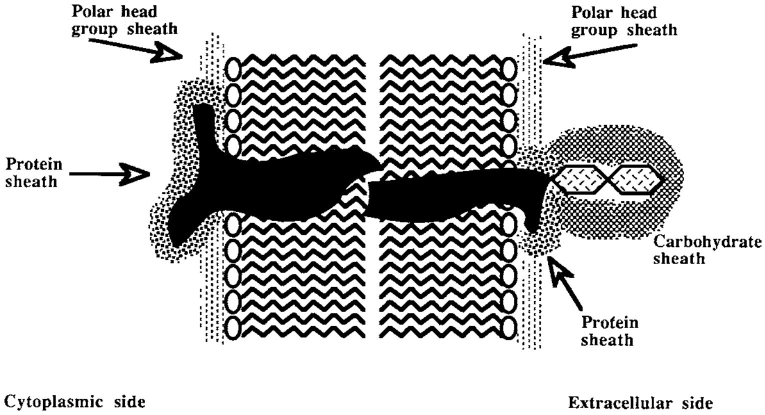


Fig. 24.6 Regions traversed by a molecule traveling from the extracellular milieu across the cell membrane

The first two assumptions permit the use of synthetic liposomes as the model system for passive diffusion directly through the membrane lipid bilayer, whose lipids are in the crystalline or liquid crystalline state. Liposomes have very low water permeability since the free energy requirement for forcing water through the hydrophobic portion of the bilayer is high. A molecule traveling from the bulk water in the extracellular milieu will traverse six barrier regions (Fig. 24.6):

1) The solvation layer surrounding the carbohydrate ends of the membrane proteins.
2) The solvation layer surrounding the proteins themselves.

3) The solvation layer surrounding the polar head groups of the outer-layer phospholipids.
4) The non-aqueous lipid layer consisting of the fatty acid components of the outer and inner layers.
5) The solvation layer surrounding the polar head groups of the inner-layer phospholipids.
6) The solvation layer surrounding the proteins on the inner cytoplasmic face of the membrane.

The activation energy for this process through a phospholipid bilayer is approximately 37 kJ/mol. The process is thought to be facilitated by oscillations of the membrane structure, which create localized "holes" into which a water molecule can insert itself. The water permeation process is facilitated at higher temperatures where there is higher kinetic energy and a higher degree of kinking or *cis* configuration formation. For a long time this passive mechanism was thought to be the prime mechanism for water transmission through biological membranes. While certainly a part of the story, the passive mechanism has been supplanted by evidence for water specific pores.

For a mammalian cell membrane containing embedded proteins, the activation energy has been found to be approximately 12 kJ/mol. If the cell is treated with *p*-chloromercury benzoate or another –SH group blocking agent, the activation energy again becomes approximately 37 kJ/mol. This suggested that the easier water permeation through normal cell membranes is attributable to pores or channels that are –SH-reagent sensitive. The family of pores that control water flow through membranes are now well known and are called *aquaporins*. Aquaporins are responsible for the rapid single file transport of water driven by osmotic force. Importantly, aquaporins do not permit the passage of charged species thus maintaining electroneutrality across the cell. The internal electrostatic field within the channel acts to limit movement of any molecule except water and is a good example of how protein structure can create local electrostatic potentials that control flux across a potential energy surface.

A dynamic equilibrium between a cell and its surroundings can be achieved, provided there are enough energy stores within the cell to maintain that equilibrium. Under such conditions, the cell volume remains constant, ATP is produced metabolically at a low maintenance rate, and the actual passive influx of Na^+ and efflux of K^+ ions is counteracted by the ATPase-driven pumps. As the energy stores of the cell are used up, ATP can no longer be adequately replenished. As a result, the Na^+/K^+ concentration ratio will be perturbed, the transmembrane potential will drop (become less negative), and the cell will swell as more H_2O, co-transported with passive Na^+ influx, enters. In the actual cell, the osmotic pressure and the transmembrane electrical potential act in concert with the chemical potentials of all the components to establish an equilibrium, that is, to establish conditions under which $\Delta G = 0$. However, the cell does not remain static, nor does it necessarily maintain the same values of the individual terms, even though overall equilibrium has been established.

As discussed earlier, the aqueous molecules will exist in a number of phases:

1) As bulk H_2O.
2) As H_2O in the outer and in the inner Helmholtz layers solvating the phospholipid head groups, including both charged and uncharged polar species.
3) As H_2O associating with other aspects of microscopic membrane structure, including outer membrane surface proteins and glycoproteins.
4) As H_2O within the lipid bilayer (e.g., in a channel), whether free or bound to a co-transported molecule or ion.
5) As H_2O dipoles oriented in the inner and outer Helmholtz layers on the cytoplasmic side of the membrane.
6) As H_2O interacting with the various ions within the cytoplasm.
7) As H_2O solvating the cytoplasmic proteins.
8) As H_2O interacting with granule or other organelle membranes, etc.

It is unlikely that H_2O acts ideally in any of these states, with the possible exception of the bulk external H_2O where its activity coefficient will be approximately 1. The "bulk" H_2O in the cytoplasm of an erythrocyte can have an activity coefficient as low as 0.2. In the red blood cell, water constitutes less than 25% by weight of the cell, and little truly free pentagonal/hexagonal hydrogen-bonded bulk water is likely to exist. The complete picture of the behavior and organization of water in a cellular system is quite clearly very complex and is not completely known at the present time.

Further Reading

General

Heinz E. (1981) *Electrical Potentials in Biological Membrane Transport*. Springer-Verlag, New York.

Hille B. (2001) *Ionic Channels of Excitable Membranes,* 3rd edition. Sinauer Associates, Sunderland, MA.

Israelachvili J. (1992) *Intermolecular and Surface Forces*, 2nd edition. Academic Press, London.

Kotyk A., Janacek K., and Koryta J. (1988) *Biophysical Chemistry of Membrane Functions*. Wiley, New York.

Specific Articles

Cafiso D., McLaughlin A., McLaughlin S., and Winiski A. (1989) Measuring electrostatic potentials adjacent to membranes, *Methods Enzymol.,* **171**:342–364.

Itoh S. and Nishimura M (1986) Rate of redox reactions related to surface potential and other surface-related parameters in biological membranes, *Methods Enzymol.,* **125**:58–86.

Jordan P.C. (1986) Ion channel electrostatics and the shapes of channel proteins, In Miller C. (ed.) *Ion Channel Reconstitution*. Plenum Press, New York, pp. 37–54.

Krämer R. (1989) Modulation of membrane protein function by surface potential, *Methods Enzymol.*, **171**:387–394.

McLaughlin S. (1977) Electrostatic potentials at membrane-solution interfaces, *Curr. Topics Membr. Transp.*, **9**:71–144.

McLaughlin S. (1989) The electrostatic properties of membranes, *Ann. Rev. Biophys. Biophys. Chem.*, **18**:113–136.

Neher E. and Sakmann B. (1992) The patch clamp technique, *Sci. Am.*, **266, 3**:44–51.

Stühner W. (1991) Structure-function studies of voltage-gated ion channels, *Annu. Rev. Biophys. Biophys. Chem.*, **20**:65–78.

Aquaporin and Water Permeation Through Membranes

Agre P. (2006). The aquaporin water channels. *Proc. Am. Thorac. Soc.*, **3**:5–13.

Agre P. and Kozono D. (2003) Aquaporin water channels: molecular mechanisms for human diseases. *FEBS Lett.*, **555**:72–78.

Schrier R.W. (2007). Aquaporin-related disorders of water homeostasis. *Drug News Perspect.*, **20**:447–453.

Problem Sets

1. The voltage change across the plasma membrane in a firing neuron is –80 mV → +50 mV.

   (a) What is the change in electric field when the neuron fires? The membrane is 6 nm thick.

   (b) The capacitance of the membrane is 1×10^{-6} F. How many charges pass through the membrane to cause the measured voltage change?

Chapter 25
Kinetics – Chemical Kinetics

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_25, © Springer Science+Business Media, LLC 2010

25.1 The Equilibrium State Is Found by Chemical Thermodynamics but Chemical Kinetics Tells the Story of Getting There

Throughout much of this book we have emphasized the description of a system in terms of an equilibrium state. This approach is invaluable when we are beginning study and are not sure exactly what are the elements and the linkages that comprise the rules of the system under study. Still it seems to be a compulsive pulse taking that is stodgy. We really want to get to the visceral issue of our curiosity; what are the details of how it works: what are the wheels and gears of the system? That question is not always easily answered and as we will see in the following section is often approached using the thermodynamic tools we have already discussed. After all, knowing how the wheels and gears are arranged is invaluable when trying to understand how a machine works. However, the joy in the thing really comes when you watch it in motion. For example, it is almost impossible to appreciate how the gears, valves, lifters, and cylinders in an internal combustion engine work without watching the thing go around. Watching a system move between states and discovering the path that it takes to get there is the job of kinetics. Given the example of the engine above it seems likely that thermodynamic and kinetic analysis will be connected in some integrated fashion. We will find this to be true. Our task then is to learn how we can capture the motion of a system's machinery and learn from these motions: how does it do that?

In chemical equilibrium-thermodynamics or chemical statics the essential questions are

"Where is a reaction going?" and,
"What are the energetic costs associated with it getting there?"

The corollaries to these central questions are

"Is the reaction spontaneous?" and,
"How far will it go or what will be the yield?"

The practical among us might add

"What can we do to increase the yield or decrease the energy costs?"

Chemical kinetics asks

"How fast is the reaction getting to where its going?"
"What is the reaction path for the reaction?", and
"What are the rates of each of those steps?"

The corollary to this question is not actually a kinetics question but is probably the focus of our interests:

> "What is the mechanism by which each step of the reaction occurs?"

Again, the more practical ones among us might add

> "Can we alter the rates of the reaction?", and if so,
> "at which step is control best applied (both for maximal effect and minimal cost)?"

25.2 A Historical Perspective on the Development of Chemical Kinetics

Chemical kinetics has biochemical origins. In 1850, L. Wilhelmy studied the inversion of an acidic aqueous solution of sucrose into its constituents, glucose and fructose.

$$H_2O + \underset{\text{sucrose}}{C_{12}H_{22}O_{12}} \rightarrow \underset{\text{glucose}}{C_6H_{12}O_6} + \underset{\text{fructose}}{C_6H_{12}O_6} \tag{25.1}$$

This process is important in the food industry as well as in home-made jelly and jam making. It results in an inexpensive but deliciously sweet mixture called invert sugar. Invert sugar is sweeter than sucrose or cane sugar itself because fructose is nearly twice as sweet as sucrose. The chemical composition of invert sugar, 1:1 glucose:fructose, is identical to the composition of honey and thus invert sugar is a cheaper and ersatz honey. Sucrose is easily hydrolyzed in acid solution and it was this reaction that Wilhelmy was studying (Fig. 25.1). Each of the saccharide components of this system is optically active with sucrose being the most dextrorotary ($[\alpha]_D^{20} = +66.5°$). The optical activity of D-glucose ($[\alpha]_D^{20} = +52.5°$) and D-fructose ($[\alpha]_D^{20} = -92°$) is substantially more levorotatory. An equimolar mixture of D-glucose and D-fructose is levorotatory. As the hydrolysis proceeded, Wilhelmy was able to record the decreasing concentration of sucrose by following the loss of dextrorotation of light with polarimetry. Wilhelmy discovered that the rate of decrease in the concentration of sucrose (c) with respect to time (t) was proportional to the concentration of the sucrose that remained unconverted:

$$-\frac{dc}{dt} = k_1 c \tag{25.2}$$

The constant $\boldsymbol{k}_1$ is called the *rate constant* of the reaction and was found to be proportional to the concentration of the acid in the mixture. The acid itself was not consumed in the reaction and so acted as a catalyst increasing the rate of the reaction without itself being changed. Integration of Eq. (25.2) gave the following expression:

$$\ln c = -k_1 t + C \tag{25.3}$$

Wilhelmy then noted that at the start of the experiment, $t = 0$, the concentration of $c = c_0$ which allowed him to evaluate C:

$$\ln c_0 = -k_1 t(0) + C \tag{25.4}$$

$$C = \ln c_0 \tag{25.5}$$

He could then substitute Eq. (25.5) into (25.3) and algebraically manipulate to get an equation that would yield $c(t)$ for any t given the initial concentration of sucrose and a rate constant; Eq. (25.7):

$$\ln c = -k_1 t + \ln c_0 \tag{25.6}$$

$$c = c_0 e^{-k/t} \tag{25.7}$$

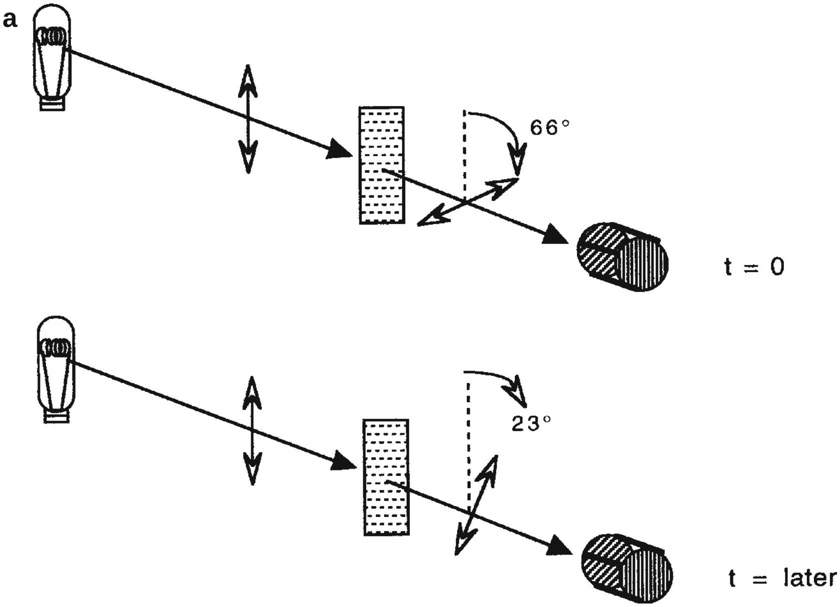


Fig. 25.1 Experimental setup for Wilhelmy's study of the inversion of sucrose. (**a**) Polarimetric analysis of the inversion of the optical rotation of plane polarized light. (**b**) The change in sucrose concentration with time as measured by changing rotation of light. The exponential data from the reaction (*plot left*) can be linearized by plotting In$\frac{[S]}{[S_o]}$against time (*right plot*). This gives a rate constant k that in this example is 3.6×10^{-3} min^{-1}

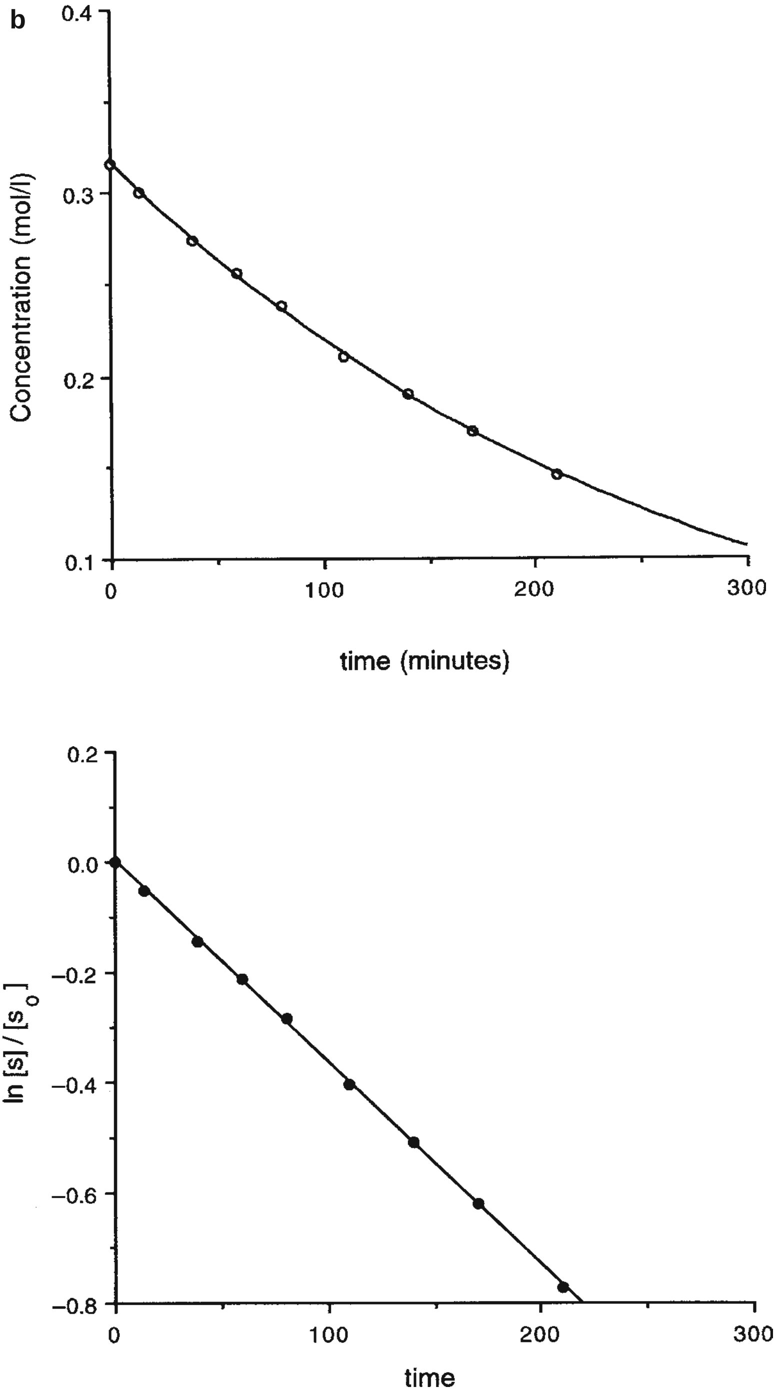


Fig. 25.1 (continued)

His experimental results closely followed this equation showing an exponential decrease in dextrorotatory power of the solution with time. Thus the ideas of chemical kinetics were introduced and experimentally validated.

The connection of kinetics to chemical equilibrium was made over the next several decades through the ideas of *dynamic equilibrium* and the *law of mass action*. A.W. Williamson noted as early as 1850 that equilibrium had a dynamic character, the result of two opposite changes occurring at the same rate. In 1862, M Berthelot and P. de St. Gilles had demonstrated, by a study of the hydrolysis of esters, that by varying the concentrations of the reactants, the concentration of the product could be varied. This work made the observation that "the amount of ester formed in each moment is proportional to the product of the reacting substances and inversely proportional to the volume." They deduced the expressions for the velocity of the reaction but did not generalize their observations into the law of mass action because they neglected the rate of the reverse reaction in their treatment. In the following year, P. Waage and C.M. Guldberg, a chemist and applied mathematician, respectively, also recognized that a system in equilibrium was not characterized by the cessation of all activity but rather was a dynamic system. They conceived the chemical process in mechanical terms. To their thinking equilibrium was a state in which a force k, was driving the reactants toward products, and a second force of equal but opposite direction was also driving products toward reactants.

> When two substances A and B are transformed by double substitution into two new substances A′ and B′, and under the same conditions A′ and B′ can transform themselves in A and B. . . The force which causes the formation of A′ and B′ increases proportionally to the affinity coefficients of the reaction A + B = A′ + B′ but depends also on the masses [concentration] of A and B. We have learned from our experiments that the force is proportional to the product of the active masses of the two substances A and B. If we designate the active masses of A and B by p and q, and the affinity coefficient by k, the force = k.p.q. Let the active masses of A′ and B′ be p′ and q′ and the affinity coefficient of the reaction A′ + B′ = A + B be k′. This force is in equilibrium with the first force and consequently kpq = k′p′q′. By determining experimentally the active masses p, q, p′, q′, we can find the ratio between the coefficients k and k′. On the other hand, if we have found this ratio $\frac{\mathrm{k}}{\mathrm{k}'}$ we can calculate the result of the reaction for any original condition of the four substances.

In 1877 J.H. van't Hoff modified the somewhat nebulous idea of "force" with the "velocity of the reaction." The velocity of $A' + B' = A + B$ is $v = kpq$ and the velocity of $A' + B' = A + B$ is $v = k'p'q'$. The velocity of the complete reaction is $V = v - v'$ $= kpq - k'p'q'$. At equilibrium, $V = 0$ and $\frac{p'q'}{pq} = \frac{k}{k'}$. Then in 1889, S. Arrhenius concluded that molecules must achieve an activated state before they could react (i.e., the molecule must reach a higher energy level). Based on earlier work by van't Hoff in which the equilibrium constant was derived from the energy of reaction, absolute temperature, and the gas constant, Arrhenius was able to formulate the relationship for the rate constant of a reaction that included the activation energy: $\ln k = -\frac{E_{\mathrm{a}}}{RT} + \text{constant}$.

25.3 Kinetics Has a Specific and Systemic Language

Kinetics, like thermodynamics, has its own particular usage of language which needs to be understood. When discussing a reaction under study, there are molecules that undergo change, the *reactants*, *products,* and *intermediates*. These can be identified because their concentrations change in the course of a reaction. The concentration of reactants decreases while that of products increases. Intermediates are formed in a multi-step reaction first as a product of one reaction and subsequently becoming a reactant in a following reaction. The concentration of intermediates will often increase and then decrease as the reaction proceeds.

There is also a class of molecules that participates in the reaction but whose concentrations do not change over the time of the reaction. These include the *solvent* and *general physical environment*, which are usually regarded to be in such excess that their effective concentrations do not change. Certain components are buffered because they are in equilibrium with a large reservoir and there is no apparent change in their concentration, this may include intermediates in a steady-state process. Finally *catalysts*, which affect the energy of activation, can speed up or slow down a reaction but emerge unchanged from the reaction. Catalysts commonly include *enzymes* and *active surfaces*.

25.3.1 Mechanism and Order

The equilibrium constant and its expression in terms of concentrations and ratios of reactants and products, $K = \dfrac{[C]^c\,[D]^d}{[A]^a\,[B]^b}$, reflect the stoichiometry of a reaction and is the expression of the law of mass action. It is important to emphasize that the stoichiometry reflects the proportion of reactants and products. There are an infinite number of sets of coefficients that can be written that will satisfy the law of mass action expression. While we already know that the link between this equilibrium expression and the dynamic path that leads to a dynamic equilibrium is expressed in kinetic terms, i.e., the sum of the rate constants, the stoichiometry itself does not help us determine the path or the *mechanism* of the reaction. The reaction mechanism is defined as the set of elementary chemical equations that define the path followed by the reactants as they are transformed into the products. The mechanism must be consistent with the stoichiometry but in general many paths can be proposed that lead from reactants to products. Under a particular set of conditions, one path may be favored over the other and this path can only be found by experimentation. If we consider the series of elementary reactions that constitute a proposed mechanism, it should be obvious that each of these equations will have its own rate expression. The sum of the coefficients of these rate expressions is called the *order* of the reaction and relates the velocity of the overall reaction to the stoichiometric concentrations.

25.4 Order of a Reaction Relates the Concentration of Reactants to the Reaction Velocity

The reaction $A + B \rightarrow C$ has a rate expression of the form

$$v = kc_A^k c_B^l c_C^m \tag{25.8}$$

where the concentrations of each of the stoichiometric components, *A, B, C*, ..., are raised to some power, *k, l, m*, ... These exponents may take any value though integer values and 0 are most commonly found. *It is emphasized that these exponents bear no relation to the stoichiometric ratios*. Zero is a particularly important exponent; it means that the velocity of the reaction, with respect to the concentration of that reactant, is constant during the reaction. A zero-order reaction is seen when the rate is limited by the concentration of a catalyst or in a photochemical reaction in which the rate is determined by the light intensity. This important special case will be of value when we consider enzyme kinetics. Several points should be highlighted:

(1) The order of the reaction with respect to each component can be described and is equal to the coefficient, *k, l, m,* of that component.
(2) The order of the overall reaction is equal to the sum of each of the exponents in Eq. (25.8).
(3) The order of the overall reaction is not deducible from the stoichiometric equation. Like the mechanism of the reaction, the order must be found experimentally.
(4) Since the order of the reaction has many of the qualities of the mechanism, it is natural to suppose that both are related to the path by which reactants become products. Studying the kinetic order of a reaction is one way we can capture the system in motion and deduce how the gears and wheels intermesh.

The order, which connects reaction velocity with concentration, is a reflection of the actual workings of a reaction but can be quite indirect. The order reflects the set of elementary equations and their rate laws. It is useful to help select the correct mechanism. However, the order of each elementary reaction must be differentiated from that of the overall reaction. This is important since the order of the overall reaction often depends on specific conditions which may affect the apparent role of each component. We will consider two examples.

The first case is to consider the order of the inversion reaction of sucrose as performed by Wilhelmy. The order of the overall reaction as performed by Wilhelmy was one, dependent only on the concentration of sucrose. The rate law for this reaction is generally written:

$$v = k\,[\text{sucrose}] \tag{25.9}$$

However, the conditions chosen by Wilhelmy were an excess of acid and an aqueous solvent. The real rate law is

$$v = k\,[\text{sucrose}][\text{H}_2\text{O}][\text{H}^+] \tag{25.10}$$

Experimentally, it can be demonstrated that the inversion of sucrose is first order with respect to [H^+] as well as [H_2O]. They are left out of the overall rate law when they are in excess because under those conditions their concentrations are constant and hence they become 0 order with no impact on the overall order of the reaction. If the reaction is run under non-aqueous conditions in which water was added as a reactant, the order of the rate law would increase to reflect this. Even though the acid remains unchanged in the reaction and is in fact a catalyst, if the acid were not in clear excess, the order of the reaction would rise to reflect the participation of the acid in the overall mechanism of the reaction. Thus while the order is a window into the running machinery of the system, it is often important to look through that window from a variety of angles to capture all of the information available from that portal.

The second case to consider is how manipulating the order of an equation can allow us to study certain aspects of the reaction. The component under study may not even be one of the reactants or products. If the reactants are all in excess such that their concentration is unchanged during the course of the reaction then the velocity of the reaction will be constant and independent of the concentration. This is the characteristic of a zero-order reaction and is also the characteristic of reactions catalyzed by enzymes when they are substrate saturated (i.e., the reaction is run with an excess of reactants). Most enzymes are first order at sub-saturated values and become zero order when saturated. We will explore in some detail how we can take advantage of these kinetic properties to learn about the properties and mechanisms of enzymes.

Given the order what does it actually tell us about the mechanism? The answer is often frustratingly little. But order is measurable from experiment and it sometimes provides a tangible constraint on the possible elementary reactions that can be proposed for an overall reaction.

25.5 Expressions of the Rate Laws Are Important Properties of a Reaction

25.5.1 Zero Order Reactions

The velocity of a zeroth order reaction is constant and is independent of the reactant concentration. The rate expression of a zero order reaction is

$$-\frac{d\,[A]}{dt} = k_{\text{o}} \tag{25.11}$$

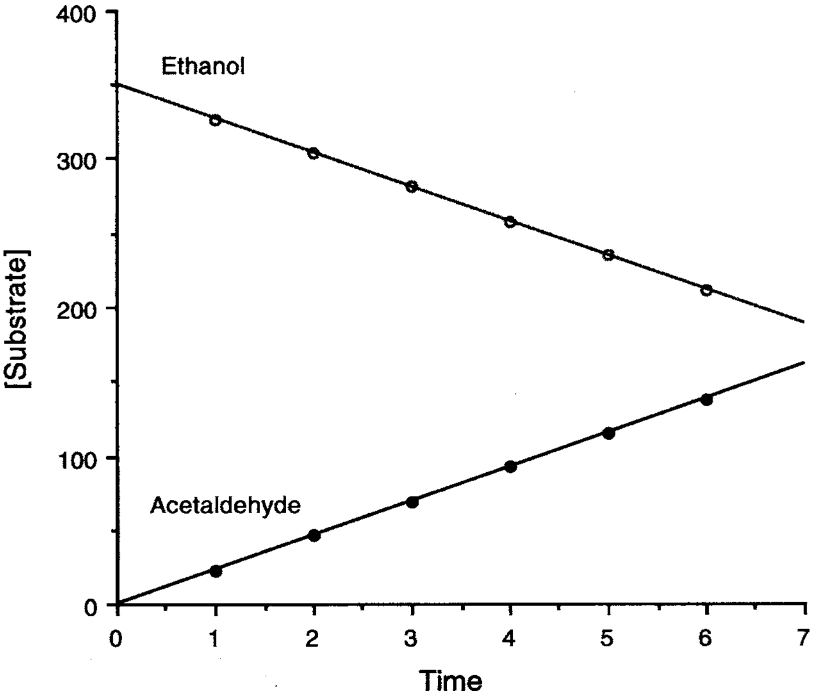


Fig. 25.2 Time course of a zero order reaction. The reaction for the conversion of ethanol to acetaldehyde catalyzed by the enzyme alcohol dehydrogenase is shown in terms of the experimentally measured linear loss of ethanol with time accompanied by the linear increase in concentration of acetaldehyde. The velocity of the reaction is the slope of the line, $v = k_o$. The course of this reaction is typical for enzyme-catalyzed reactions when there is excess substrate. As the concentration of ethanol falls during the course of the reaction, the reaction order will change and no longer be zero-order

and its progress as a function of time can be seen in Fig. 25.2. It is hard to imagine what the molecularity of an elementary reaction demonstrating zero-order kinetics would look like since the reaction apparently proceeds without respect to the molecules involved. However, if the rate is limited by the concentration of a catalyst the reaction will appear to be zero order. Photochemical reactions in which the reaction depends on light intensity will also be described by these kinetics. Upon integration of this differential equation, the following expression is obtained:

$$[A_o] - [A] = kt \tag{25.12}$$

with units of mol/L/s.

25.5.2 First-Order Reactions

The velocity of a first-order reaction varies with the concentration of a single reactant. The rate expression and the integrated rate law for a first-order reaction has already been described and illustrated in the text but is summarized here. For reaction

$$\mathrm{A} \rightarrow \text{products} \tag{25.13}$$

the rate equation is

$$-\frac{d\,[A]}{dt} = k\,[\mathrm{A}] \tag{25.14}$$

After integration with the concentrations of A at $t_1 = 0$ and t_2

$$\ln \frac{[A]_0}{[A]} = kt \tag{25.15}$$

which can be written

$$[A] = [A]_0\, e^{-kt} \tag{25.16}$$

or

$$\ln [A] = \ln [A]_0 - kt \tag{25.17}$$

Equation (25.17) can be employed to provide a useful parameter of kinetic processes, the *half-life*, $t_{1/2}$. The half-life is the time necessary for 1/2 of the original substance to react. Since $\ln \left[\frac{1}{2}\right]$ = 0.693, the half-life is given by 0.693 τ where τ is the time constant for the reaction $\left(\tau = \frac{1}{k}\right)$. The half-life of a first-order reaction is independent of the initial concentration of the reactant thus the time constant for a reaction is the same regardless of the sample size. Thus first-order reactions behave like the independent Bernoulli trials summarized in Chapter 5. An important example of first-order kinetics and the half-life effect is seen in the radioactive decay of nuclei. Radioactive decay occurs with the emission of α (He nuclei), or β (electrons) particles or γ photons. The decay of a particular radioactive nucleus is independent of any other radioactive nucleus and thus gives rise to the effect that one half of the radioactive material at t=0 has decayed at $t = t_{1/2}$. At $2t_{1/2}$, $\frac{1}{4}$ of the original material will remain radioactive and so on.

25.5.3 Second-Order Reactions

The velocity of a second-order reaction will vary with the square of the concentration of a single reactant or with the concentration of two reactants. It is impossible to determine which mechanism of these is operating on the basis of the order alone. For the first of these reactions

$$A \rightarrow \text{products} \tag{25.18}$$

and for the reaction

$$A + B \rightarrow \text{products} \tag{25.19}$$

where $[A] = [B]$ at $t = 0$, the rate equation is

$$-\frac{d\,[A]}{dt} = \mathrm{k}\ \ [\mathrm{A}]^2 \tag{25.20}$$

The integration of

$$-\frac{d\,[A]}{[A]^2} = kdt \tag{25.21}$$

yields

$$kt = \frac{1}{[A]} - \frac{1}{[A]_0} \tag{25.22}$$

The second-order rate constant can be determined from the slope of the straight line generated by a plot of $\frac{1}{[A]}$ against t. The half-life of this reaction is given by

$$t_{1/2} = \frac{1}{k\,[A]_0} \tag{25.23}$$

If the reaction under study is

$$aA \rightarrow \text{products} \tag{25.24}$$

then wherever the rate constant appears, ak must be substituted. If the reaction is

$$aA + bB \rightarrow \text{products} \tag{25.25}$$

and $b[A]_0 \neq a[B]_0$ then

$$kt = \frac{1}{b\,[A]_0 - a\,[B]_0} \ln \frac{[A]\,[B]_0}{[A]_0\,[B]} \tag{25.26}$$

The rate constants have the units of L/mol/s.

25.5.4 Experimental Determination of a Rate Law Requires Measurement of Two Observables, Time and Concentration

Concentration and time are the observables from which the rate law for a reaction is derived. The most straightforward analysis is simply to stop a reaction at a particular time point (or series of time points) following $t = 0$ and to measure the concentrations of the reactants, products, intermediates, and presumed catalysts. This may seem straightforward but in practice it is often an exceedingly difficult task to approach kinetic experiments in this fashion. It is far more preferable to allow a reaction to proceed to completion without interference and to obtain the required concentration changes through some physical parameter such as spectral absorbance or optical rotatory power which can simultaneously be followed without interfering with the running of the reaction.

Among the various kinetic tools available is the *method of isolation* in which all but one of the reactants is present in very high quantities. Then the reaction will have the order of only the low-concentration reactant. If, for example, the order of reaction of this component is first order, the overall reaction will be called pseudo-first-order. Another method of obtaining the data is to stop a fast reaction by sudden cooling or denaturation (especially useful in enzyme kinetic studies). This can be performed in a stop-flow apparatus in which the quenching agent is added to a second mixing chamber that is in line after the initial reactants were suddenly added together in a first mixing chamber. The flow time between the two chambers can be quite short (milliseconds) so rapid reactions can be studied. For very fast reactions, transient analysis must be used in which the conditions of a reaction at equilibrium are suddenly changed and the reaction is followed as the system relaxes back to a new equilibrium. Application of these methods is discussed in greater detail in the references at the end of the chapter.

25.6 Elementary Reactions Are the Elements of the System That Defines a Chemical Mechanism

As we have already defined, the mechanism of a reaction is defined as the set of elementary reactions comprising it. The concept of elementary mechanisms has value because each of these reactions form a building block that can be understood in terms of geometry, energetics, and potential energy surfaces independent of the specific overall reaction in which it may participate. Gaining insight into an elementary reaction then is similar to the idea of interactional energies and chemical bonding which are more generalized expressions of chemical affinity than a specific molecular structure might imply. The order of a reaction, which is always determined experimentally, may not reflect all of the elementary reactions needed to complete the reaction and cannot be used to deduce the mechanism. However, the mechanism can always be used to deduce the order of a reaction. The elucidation of an elementary reaction includes

(1) a description of the type and number of molecules that participate in the reaction and,
(2) the energies of and the accessibility to the activated or transition states that are formed in the path of the elementary reaction.

The number of molecules that participate in a given elementary reaction is called its *molecularity*. Generally elementary reactions are *unimolecular*, *bimolecular,* or *trimolecular*. These reactions are first-, second-, and third-order, respectively, and the order can be deduced from knowledge of the chemical mechanism at this level. The converse is not true; the molecularity of the reaction could not be deduced from its order. Reactions that occur in solution or at surfaces are often very complex in terms of the number of molecular species that affect the energy and composition of the transition state and molecularity can become a somewhat abstruse concept. In gases or in ideal solutions the concept of molecularity is more useful. It is worthwhile none-the-less to have a feel for the language of molecularity. Included in the idea of molecularity lie the theories of the activated state and the transition state which will complete our theoretical treatment of kinetics.

25.7 Reaction Mechanisms Are a System of Interacting Elements (Molecules) in the Context of a Potential Energy Surface

The fundamental idea of reaction mechanics is that the reacting molecular species acquire energy which allows an *excited* state to be formed. The excited state is a state of higher potential energy molecule in which the molecule is considered to be *activated*. If more than one molecule is needed to be placed in the activated state an *activated complex* is said to exist. Molecularity can be defined as the number of molecules needed to form an activated complex. If we consider the potential energy state of the reactants, activated state, and products on a potential energy surface, the reactants and products each occupy local minima and the activated state lies at a saddle point that separates the minima wells (Fig. 25.3). The saddle point is a local maxima and thus the activated state lies at the top of a potential energy barrier. The energy required to go from a local minimum to the activated state is the *activation energy*. The path along the potential energy surface represents the reaction mechanism.

25.7.1 Collision Theory

We consider in simplest terms a bimolecular reaction in which an activated complex is formed when two molecules each possessing a certain kinetic energy collide and shatter into new particles or products. This *collision theory* is consistent with the

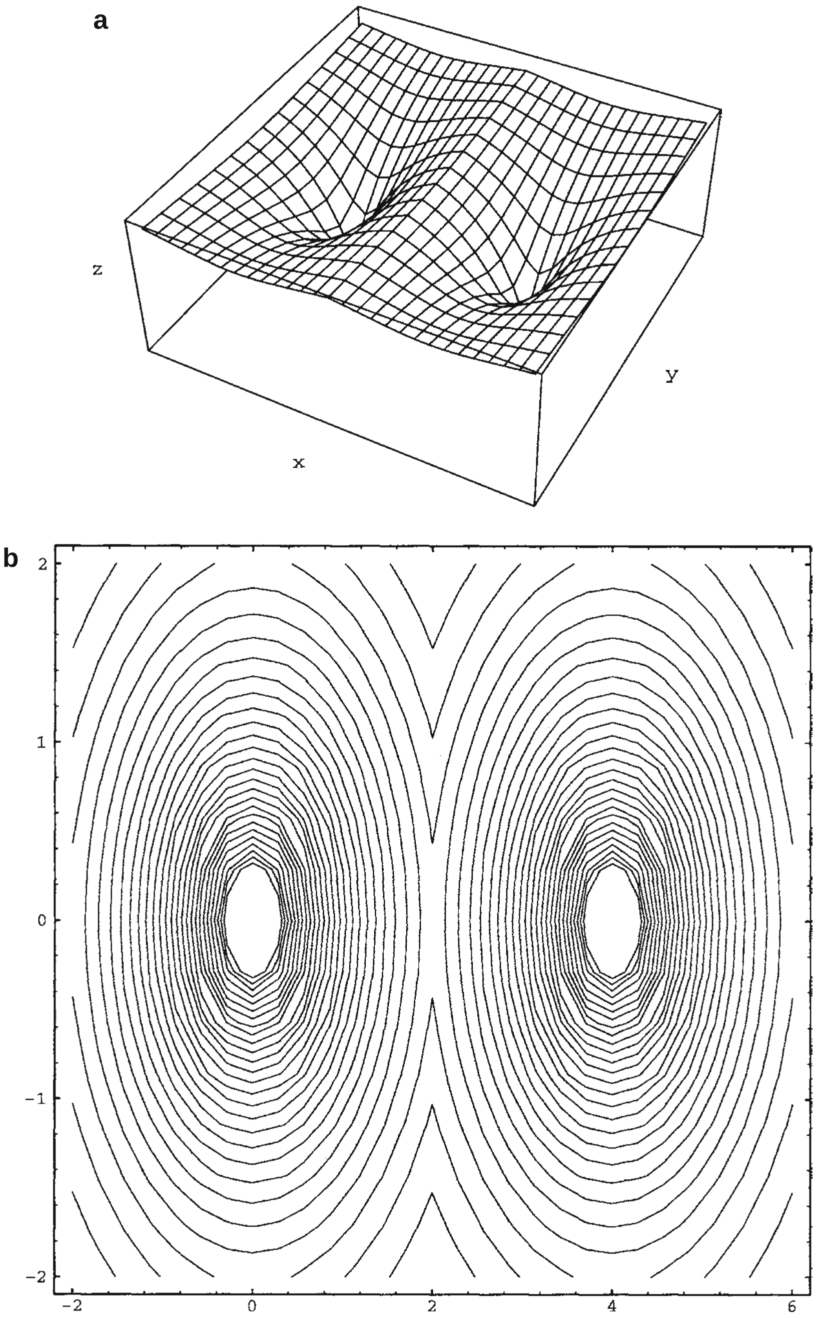


Fig. 25.3 Representations of the potential energy surface in a chemical reaction. (**a**) A three-dimensional surface showing the minimum wells and the saddle point, (**b**) a contour plot representation of the three-dimensional surface, and (**c**) a two-dimensional reaction coordinate diagram showing the thermodynamic states of the reactant, activated and product states, and the transition energy

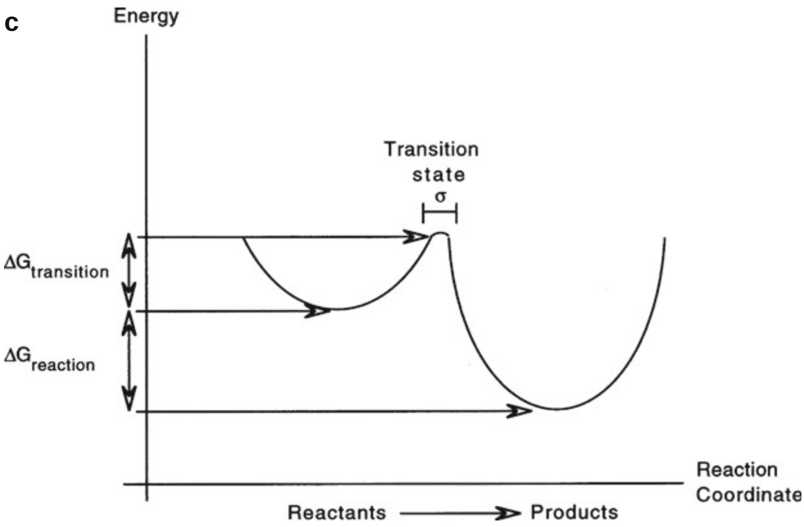


Fig. 25.3 (continued)

observations of Arrhenius who noted that reaction rates were dependent on temperature such that the natural logarithm of the rate constant for a reaction, ln *k*, varied in a linear fashion with $\frac{1}{T}$. A plot of ln *k* against $\frac{1}{T}$ gave a straight line that was a constant characteristic for a given reaction as shown in Fig. 25.4. The mathematical form of this observation is called the *Arrhenius equation*:

$$\ln k = \ln A - \frac{E_\mathrm{a}}{RT} \tag{25.27}$$

or equivalently

$$k = Ae^{-E_\mathrm{a}/RT} \tag{25.28}$$

The pre-exponential parameters *A* and E_a, which is the activation energy, are the *Arrhenius parameters* for a reaction. The units of *A* are the same as those of *k* (i.e., A/s^{-1} for a first-order and A/L mol^{-1}s^{-1} for a second-order reaction). The activation energy is given in kJ/mol^{1}. The temperature dependence of the rate can be found by comparing the rate constants at two different temperature using Eq. (25.27) or (25.28). The collision theory is appreciated in comparison to two hard spheres such as billiard balls that collide. If a sufficient kinetic energy is present when the balls collide, they will break and form new particles or products. If, on the other hand, not enough energy to create the new product state is available, on collision they will bounce apart without undergoing a state change. The energy necessary to cause the products to form is the activation energy and sufficient energy must be available to surmount the activation barrier and form new products. The Arrhenius equation and the collision model are most easily imagined with respect to reactions in a gas.

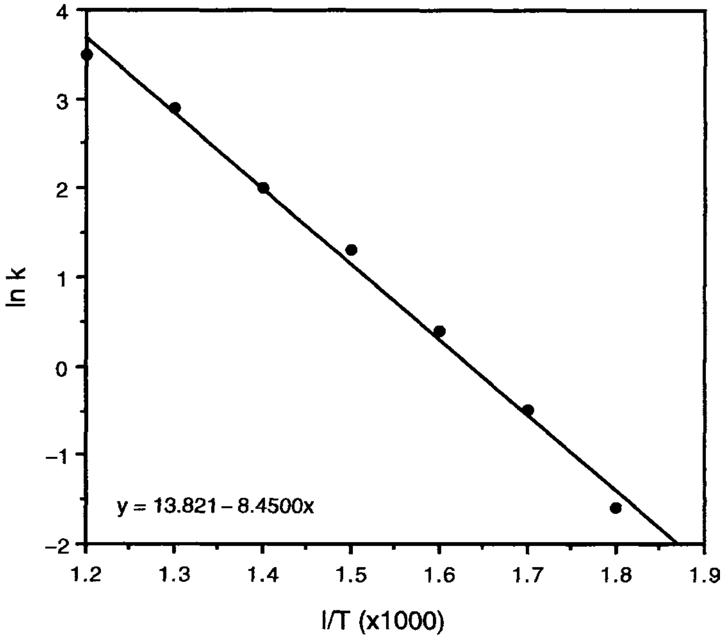


Fig. 25.4 Experimental relationship leading to the Arrhenius equation. The ln k is plotted on the ordinate and $\frac{1}{T}$ on the abscissa. The slope of the line is equal to $-\frac{E_a}{RT}$ and the y-intercept is ln A. This line describes an equation of the slope intercept form: ln k = intercept + slope $\times$ $\frac{1}{\mathrm{T}}$. The terms A and $\frac{E_a}{RT}$ are the Arrhenius parameters and can be explained in terms of collision theory

This gaseous model can be considered as follows: The rate of collisions between two reactants will be proportional to the concentrations of the molecules present. Doubling the concentration of either of these reactants will double the rate of collisions. The rate of collision is proportional to $[A][B]$. Assuming for a moment that every collision has the orientation that would allow progression to product, if adequate energy is available to surmount the activation barrier, we need to have a measure of the fraction of A and B that possess the required energy E_a. Such a measure of the fraction of a population that possesses a certain energy, f, is provided by the Boltzmann distribution and has the form:

$$f \propto e^{-E_a/RT} \tag{25.29}$$

The reaction rate will be proportional to the number of collisions times the fraction of successful collisions (only limited by energetic issues at this point):

$$v = [A]\,[B]\,e^{-E_a/RT} \tag{25.30}$$

For a second-order reaction the rate law is

$$v = k\,[A]\,[B] \tag{25.31}$$

which makes

$$k \propto e^{-E_{\mathrm{a}}/RT} \tag{25.32}$$

or if we let the constant of proportionality be A

$$k = Ae^{-E_{\mathrm{a}}/RT} \tag{25.33}$$

This demonstrates the derivation of the Arrhenius expression from the collision theory. The activation energy is the minimum energy necessary in a collision to generate products and the pre-exponential factor is a constant of proportionality that connects the concentrations and collision rate of the reactants. We assumed at the outset of this derivation that the energetics of the collision were the only limiting factor. Experimentally this is not found to be the case and the experimental values of A are often less than the calculated values from the kinetic theory of gases. This deviation is caused by the requirement of proper orientation between the colliding particles, a so-called *steric factor* P, which causes certain orientations to be non-reactive. P usually is between 0 (no orientation leading to reaction) and 1 (all orientations lead to reaction). The steric factor is an empirical correction factor that cannot be determined from any simple theory. The problem with these correction factors and with the overall treatment of bimolecular mechanisms by the collision theory is that the change in electronic structure that take place in all chemical reactions are ignored and hence the influence of these electronic redistributions on the cross-section of the colliding molecules goes unaccounted.

25.7.2 Surprises in the Collision Theory State Space Require Re-evaluation of the Abstraction

It is a reasonable first approximation to treat two neutral molecules approaching one another as a pair of hard spheres of a specific cross-sectional dimension because we know that the dominant force with very close approach is strong repulsion. This repulsion is due to the Pauli exclusion principle. This principle gives rise to the physical properties of hardness that allow us treat the molecule using the classical methods of the kinetic theory of gases. However, Pauli exclusion is a quantum mechanical property and we should anticipate some "surprise" for a classically oriented observer who is investigating molecular state space. Two "surprising" examples of collision theory will help make the case for further theoretical developments with an new abstraction that is more sensitive to electronic structure and dynamics. These mew abstractions will lead us to *transition-state theory*.

A surprising result of the collision theory is the gas-phase reaction

$$K + \mathrm{Br}_2 \rightarrow K\mathrm{Br} + \mathrm{Br} \tag{25.34}$$

The steric factor for this reaction is $P = 4.8$ which according to our earlier definition means that the reaction is occurring more frequently than the molecules are meeting! This result is absurd, *but it is our abstraction which is wrong*, not the experimental data. The molecular mechanism of this reaction appears to be the following: The K atom approaches the Br_2 molecule and at a distance considerably greater than the rigid sphere dimension, an electron is transferred from the potassium atom to the bromine molecule. The approaching collision is now between two oppositely charged ions, which move much more quickly than anticipated because of the newly generated electrostatic force. The reaction rate is substantially faster than we anticipated when our physical model was the hard sphere mechanical dimension of the two molecules. The diameter of interaction must include the electronic behavior of these molecules. This mechanism is reminiscent of our analysis of the "surprisingly" rapid conduction of aqueous protons discussed in Chapter 22.

Recognizing the presence of the electronic distributions but maintaining the hard (Pauli exclusion) sphere model, what energy is needed to deform the electronic clouds as two molecules approach collision. Two neutral molecules, i.e., with closed shells, require activation energies on the order of 80–200 kJ/mol to react in an exothermic reaction. Alternatively, the activation energy for two molecules, one neutral and one a radical (an open electronic shell), is 0–60 kJ/mol. For two radicals the activation energy is nearly 0 kJ/mol and may often be negative thus making the reaction go more slowly at higher temperature. When the reaction rates are measured, the bimolecular collisions between the neutral molecules are seldom observed if an alternative path using a lower activation energy is available. The collision theory develops difficulties with reactions like these because it assumes that the only molecular energy in involved in the activation process is the kinetic energy of translation, vibration, and rotation. There is no rational way to handle other energies such as those associated with entropy and the conformational changes that will typically be found in solutions and biological systems.

25.7.3 Transition-State Theory Is a Quantum Mechanical Extension of the Classical Flavor of Collision Theory

As these examples illustrate, as molecules approach one another, their wavefunctions overlap and there is quantum mechanical interaction long before the rigid sphere model would anticipate such association. (There is also interaction when the molecules are moving apart in this perspective.) This extended structure has the qualities of reactants and products and is in a dynamic state of electronic redistribution. This structure is the *transition state* of the reaction. The idea of a transition state is much more versatile than the collision theory which is probably over-mechanical and best suited for relatively simple reactions in the gas phase. *Transition-state theory* and its extension to dynamical modeling on a potential energy surface are useful in exploring the kinetics and chemical mechanisms in a complex biological milieu that includes liquids, surfaces, and enzymes. We will develop the ideas of absolute reaction rates using potential energy surfaces shortly but will first explore the ideas underlying transition-state theory. This theory was proposed by Eyring in 1935 with important contributions to the theory by Evans and Polyani and, Pelzer and Wigner. The basic theory has been applied to other rate processes besides chemical reactions including liquid flow, dielectric loss, the internal friction of polymers, and diffusion (cf. Chapter 22). Transition-state theory treats the movement of a system from reactant to product as a flux through a surface that is defined in coordinate space. The flux of trajectories that pass through this surface without turning back is called the *reactive flux*. For a high-potential energy barrier, this surface is obviously placed at the top of the barrier perpendicular to the reaction coordinate. Two approximations underlie transition-state theory: the Born–Oppenheimer approximation and the assumption that the Boltzmann distribution governs the arrangement of the molecules among their states.

The presence of the transition state in an elementary reaction requires that every elementary mechanism be written to explicitly include it. The bimolecular reaction

$$\mathrm{A + B} \underset{k_2}{\overset{k_1}{\rightleftharpoons}} \mathrm{C + D} \tag{25.35}$$

is rewritten in terms of the transition state as

$$\mathrm{A + B} \underset{k^{\ddagger}_{-1}}{\overset{k^{\ddagger}_{1}}{\rightleftharpoons}} \quad \mathrm{AB}^{\ddagger} \underset{k^{\ddagger}_{-2}}{\overset{k^{\ddagger}_{2}}{\rightleftharpoons}} \mathrm{C + D} \tag{25.36}$$

In similar fashion to our consideration of the collision theory we would like to derive an expression for the reaction constant. The reaction rate can be derived in terms of the properties of the reactants and the rate at which they have formed activated complexes. The rate is the concentration of activated complexes times the average frequency with which a complex moves to the product state. It is important to remember that the transition state is not an isolable intermediate but rather a structure that is falling apart either in the direction of products or backward into the direction of the reactants. It is easier to calculate the concentration of the transition state if we make the assumption that the reactants are in equilibrium with the transition state. This equilibrium state can then be treated with the thermodynamic or statistical tools that we have already discussed.

We consider a reaction path (Fig. 25.3c) along a potential energy surface such as those illustrated in Fig. 25.3a. A narrow region at the peak of the potential energy barrier is defined of arbitrary length δ. δ defines the region of the transition state and systems in this region are defined as the activated complex. We treat the case in equilibrium with the reactants and can write

$$K^{\ddagger} = \frac{\left[AB^{\ddagger}\right]}{[A]\,[B]} \tag{25.37}$$

which gives the following expression for concentration

$$\left[AB^{\ddagger}\right] = K^{\ddagger}\,[A]\,[B] \tag{25.38}$$

This expression is now written in terms of the molecular partition function per unit volume. The needed expression is derived from the following expression for the equilibrium constant:

$$K = \frac{[C]^c[D]^d}{[A]^a[B]^b} = \frac{\left(z_C/V\right)^c\left(z_D/V\right)^d}{\left(z_A/V\right)^a\left(z_B/V\right)^b}e^{-E_{\text{o}}/kT} \tag{25.39}$$

Applying this equivalence, Eq. (25.38) is rewritten as

$$\left[AB^{\ddagger}\right] = [A]\,[B]\,\frac{z^{\ddagger}}{z_A z_B}e^{-E_{\text{o}}/kT} \tag{25.40}$$

In this equation $z^{\ddagger}$ is the partition function per unit volume of the activated complex and E_0 is the height of the lowest energy level of the complex above the sum of the lowest energy levels of the reactants. As we noted above, the rate of the reaction is the concentration of the activated complex times the frequency of the passage over the barrier ($v^{\ddagger}$):

$$\frac{-d\,[A]}{dt} = \left[AB^{\ddagger}\right]v^{\ddagger} = \ k\,[A]\,[B] \tag{25.41}$$

The frequency $v^{\ddagger}$ is equal to the frequency with which a complex moves apart into products. Imagine the internal motions of the complex with vibrations along the bond axes. When one of the vibrational motions is transformed into a translational motion along the bond axis, the complex will fly apart. The direction of translation for the fragments of the separating complex is determined by the bond formerly holding the complex together. The rate constant, *k*, can be written by algebraic manipulation of Eqs. (25.38) and (25.39) which yields

$$k = v^{\ddagger}\frac{z^{\ddagger}}{z_A z_B}e^{-E_{\text{o}}/kT} \tag{25.42}$$

$\frac{z^{\ddagger}}{z_A z_B}$ is similar to a partition function for any molecule with the difference that one of the vibrational degrees of freedom is becoming translational as it passes travels along the reaction path. In Chapter 17 we wrote the partition function for a single degree of vibrational freedom:

$$z^{\ddagger}_{\text{vib}} = \frac{1}{1 - e^{-hv^{\ddagger}/kT}} \tag{25.43}$$

This vibrational motion is along the reaction path such that it becomes a translational motion and during the first such vibration, the molecule falls apart. The energy needed to shake the molecule apart is quite small and ${}^{hv^{\ddagger}}\!/_{kT}$ is much less than 1. This must be the case since given a temperature at which the reaction is observed, the vibrational mode that leads to the decomposition of the complex must, by definition, be completely excited (i.e., the bond is broken). We use this assumption to solve for the exponential by an expansion series and discard the terms beyond the first power in $\left({}^{h\nu^{\ddagger}}\!/_{kT}\right)$:

$$\begin{aligned} e^{h\nu^{\ddagger}/kT} &= 1 - \left(\frac{h\nu^{\ddagger}}{kT}\right) + \frac{1}{2}\left(\frac{h\nu^{\ddagger}}{kT}\right)^2 - \cdots \\ &= 1 - \left(\frac{h\nu^{\ddagger}}{kT}\right) \end{aligned} \tag{25.44}$$

which can be substituted into Eq. (25.43) yielding

$$z_{\text{vib}}^{\ddagger} = \left(\frac{h\nu^{\ddagger}}{kT}\right)^{-1} = \left(\frac{kT}{h\nu^{\ddagger}}\right) \tag{25.45}$$

This vibrational mode is a particular partition function of the overall partition function of the transition state and needs to be factored out of the overall function giving the following expression:

$$z_{\ddagger} = z_{\text{vib}}^{\ddagger} z^{\ddagger'} = \left(\frac{kT}{h\nu^{\ddagger}}\right) z^{\ddagger'} \tag{25.46}$$

When this expression is substituted into Eq. (25.42), the unknown frequency cancels out yielding the *Eyring equation* for the rate constant:

$$k = \left(\frac{kT}{h}\right) \frac{z^{\ddagger'}}{Z_a Z_b} e^{-E_o/kT} \tag{25.47}$$

Since the transition state is actually falling forward or backward with a certain probability, a factor κ is included with Eq. (25.46) which is the *transmission coefficient* or the probability that the transition state moves on the reaction path toward products rather than back to reactants. κ is usually assigned a value between 0.5 and 1.

$$k = \kappa \left(\frac{kT}{h}\right) \frac{z^{\ddagger'}}{Z_A Z_B} e^{-E_0/kT} \tag{25.48}$$

This theoretical rate constant contains terms that explicitly depend on the properties of the reactant molecules and the activated complex and therefore improves on the arbitrary quality of the steric factor of the collision theory.

We have discussed transition-state theory from a statistical mechanical standpoint to emphasize the dynamic quality of the transition state along with its internal motions that determine the course of the reaction path. These images are particularly helpful when exploring the role of enzymes in biochemical catalysis (vide infra). Transition-state theory can also be expressed with a thermodynamic formulation. This approach is more familiar than the partition function treatment. Recalling Eq. (25.36) and (25.37) we can write an equilibrium constant of the activated state

$$K^{\ddagger} = \frac{[AB^{\ddagger}]}{[A]\,[B]} \tag{25.37}$$

Utilizing the Eyring equation (25.47) we write for the rate constant

$$k = \left(\frac{kT}{h}\right) K^{\ddagger} \tag{25.49}$$

In Chapter 12 we explored the relationship between ΔG and K, $\Delta G^{\mathrm{o}\ddagger} = –RT \ln K^{\ddagger}$, and can express Eq. (25.49) in these terms:

$$k = \left(\frac{kT}{h}\right) e^{-\Delta G^{\ddagger}/RT} \tag{25.50}$$

We will recognize that this formulation is in the standard state but will eliminate the “o” from the following equations for clarity. ΔG is a function in terms of enthalpy and entropy and we can consider these components: the Gibbs free energy, the entropy and enthalpy of activation:

$$k = \left(\frac{kT}{h}\right) e^{\Delta S^{\ddagger}/R} e^{-\Delta H^{\ddagger}/RT} \tag{25.51}$$

The experimental entropy of activation is a good indicator of the nature of the transition state. If $\Delta S^{\ddagger} > 0$ then the entropy of the complex is greater than that of the reactants. If a transition state is tightly bound, it will have a lower entropy than one that is loosely bound. As a transition state is formed by an association between separate molecules, the translational and rotational degrees of freedom become more limited and the entropy commonly decreases (i.e., $\Delta S^{\ddagger} < 0$). If the $\Delta S^{\ddagger}$ is close to the overall ΔS for reactions of the type $A + B \rightarrow AB$, then the transition state looks very much like the product. Reactions of this type proceed surprisingly slowly as a consequence of the required fall in entropy on formation of the transition complex with its associated restricted translational and rotational modes.

25.7.4 The Potential Energy Surface Unifies the Models

We have been describing the kinetic process of change in terms of a reaction path that is traveled by reactants, transition-state entities, and products. Our analysis has shown that this path relates the internal energy of the involved species at each point along the path to the physical coordinates of the atomic structure. We know that a plot of energy with respect to molecular coordinates is a potential energy surface for a chemical species. Movement of the atomic constituents along the PES is determined by the forces acting on the atoms and these forces are the gradient of the potential energy with respect to distance:

$$F_x = -\frac{\partial V}{\partial x} \tag{25.52}$$

It is very easy to imagine how a potential energy surface is a useful abstraction for molecular conformation and behavior if a ball or collection of balls are considered to be rolling around on a contour surface with hills and depressions under the force of gravity. The inter conversion between kinetic energy (the momentum of the balls) and the gravitational potential energy (where the ball is on the surface) is easily appreciated watching the movements of the ball. We have already discussed in some detail the treatment of a molecule under the Born–Oppenheimer approximation either as a collection of balls connected by springs (a classical treatment) or as nuclei influenced by forces under quantum mechanical rules of engagement. We have also already discussed how surfaces can be created via semi-empirical quantum mechanical methods in which spectroscopic data is used versus ab initio methods. A potential energy surface is calculated by moving the nuclei to many different positions in the coordinate space and then calculating the molecular electronic energy (and wavefunction in the quantum treatments) for each of the nuclear configurations.

The concept of the transition state as a species with shared properties taken from the reactants and products allows us to easily apply our experience with potential energy surfaces to the case of kinetic behavior. We consider first a simplified and stylized molecular interaction:

$$A + BC \rightarrow AB + C \tag{25.53}$$

We know that this reaction will have a transition state

$$A + BC \rightarrow [ABC]^{\ddagger} \rightarrow AB + C \tag{25.54}$$

This reaction involves three nuclear coordinates and has three internal degrees of freedom. From our earlier discussions we remember that the number of internal degrees of freedom for a triatomic molecule is (3N – 5) for a linear molecule and (3N – 6) for a triatomic in general. We cannot easily plot the potential energy as a function of three coordinates so we fix the angle of approach of A to BC and plot the potential energy as a function of the intermolecular distance between B–C and A–B (i.e., R_{AB} and R_{BC} in Fig. 25.5). When the distance between A and B is large, R_{AB} is large and the potential energy of the system is dominated by BC. The potential energy of the BC complex is represented by the large well on the left of the diagram, which essentially extends leftward toward infinity. The potential energy of

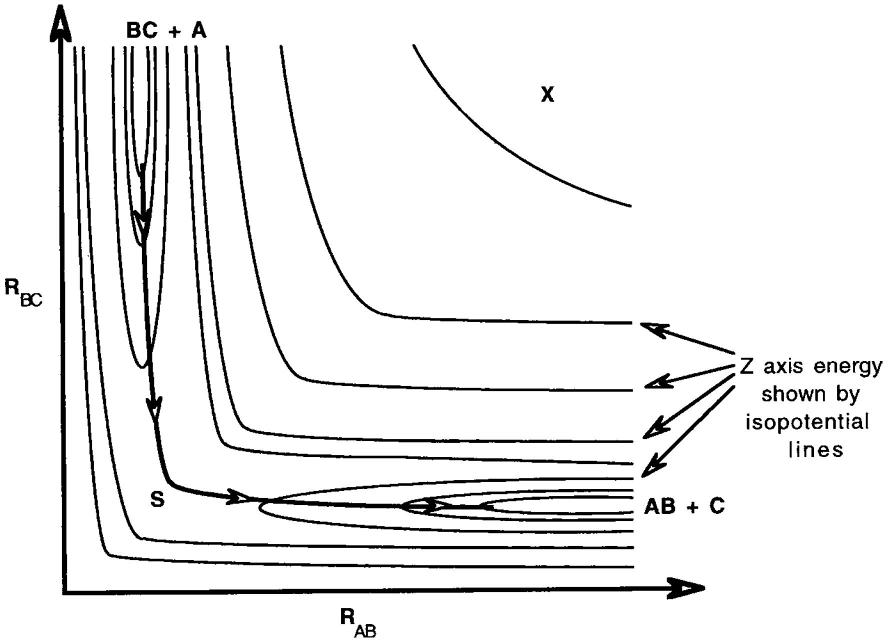


Fig. 25.5 Potential energy surface for the reaction A + BC → AB + C. The potential energy of the *BC* complex is represented by the large well on the left of the diagram, which essentially extends leftward toward infinity. The potential energy of the *AB* complex is found in the large well to the right and the potential energy of the three separate atoms is represented by the large plateau labeled *S* (From Alberty R.A. and Silbey R.J. (1992) *Physical Chemistry,* 1st edition. Wiley, New York. Adapted by permission of John Wiley & Sons, Inc.)

the *AB* complex is found in the large well to the right and the potential energy of the three separate atoms is represented by the large plateau labeled X. A reaction path of minimum energy in which *A* approaches *BC* can be drawn that shows the passage from *BC* to *AB* over a saddle point, *S*, that represents the transition state. This reaction path is often depicted on the two dimensional potential energy diagrams to show the energetics of a reaction but it is not the only possible reaction path, nor is the reaction path actually quite so straight. Let us look a little more closely at this issue of available reaction paths.

First let us consider the path as drawn. We imagine that A approaches a non-vibrating, BC, along their internuclear axis. As the distance R_{AB} decreases, the translational kinetic energy of A is converted into potential energy and A mounts the potential energy barrier at S along with BC. If enough kinetic energy is available to enable the transition-state complex to pass over the saddle point AB and C will be formed and the reaction will be observed to proceed. If insufficient energy is available, the reactants return to the potential energy valley to the left and the reaction does not proceed. This description is for reactants that are highly constrained in their motions: the approach angle is the internuclear axis, there is no vibrational motion, etc. Multiple reaction paths can be considered in which the angle of approach varies, vibrational and rotational energies vary and there are relative kinetic energies due to translational motion. The entire potential energy surface is comprised of various reaction paths each associated with different initial conditions. The actual quantum mechanical calculation of a complete potential energy surface is difficult and few reactions have been modeled in detail. A commonly cited example is the reaction of a hydrogen atom with a H_2 molecule (see Shavitt L, Stevens RM, Minn FL and Karplus M (1968) *J. Chem. Phys*. **48**:2700).

All reactions paths on a potential energy surface are possible. However, certain reaction paths are more likely than others and these will dominate the apparent mechanism for a reaction. Essentially we are describing a dynamical model with a vector field of various trajectories whose flux through the transition point is dominated by certain paths or trajectories. It is a much more convenient abstraction (in other words usually necessary) to calculate these paths using Newtonian mechanics because of the computational difficulties with the time-dependent *Schrödinger* equation. For our qualitative purposes we will limit our consideration to this Newtonian abstraction. Furthermore, we will loosen the abstraction to the level of the picturesque and imagine marbles rolling on contoured surfaces with the ability to merge and transform into and out of a transition state on their way through the saddle point. First we consider a marble, BC, that possesses a certain vibrational energy. It is approached by a marble A with translational energy. The question is whether the trajectory of BC will carry it over the saddle point to the reactant state, AB. A large number of initial states of BC and A can be picked in which different vibrational energies and modes, positions in the potential energy well, and translational energies as well as angles of interaction of both A and BC can lead to a trajectory or path. These various trajectories either will or will not carry the complex over the saddle point, i.e., the probability is either 1 or 0 that a given trajectory (which is the result of the initial conditions) will lead from reactants to product. Several of these possible trajectories are illustrated in Fig. 25.6. The rate constant is related to a series of trajectories on the potential energy surface and is a function of the probability of reaction, $P(b)$. The probability of reaction is the fraction of trajectories for a given set of reactant parameters such as the initial relative velocity that leads through a reaction cross-section to product. In other words reaction rates can be thought of as probabilities of a reaction occurring and calculated as a proportion of populated states, the reactant state versus the transition state versus the product state. It is not hard to imagine which interactions between BC and A along the potential energy surface in Fig. 25.6 might lead to reactants that will not go over the saddle point barrier. Different trajectories will approach the saddle point with varying translational or vibrational movement and may drive directly over the saddle point to product or may oscillate around the fringes of the saddle point to reach product state. Of course, many of these paths will not make it through the saddle point.

So far we have considered a potential energy surface that is symmetrical in both directions. This is not generally the case for most reactions. Some potential energy surfaces will favor a reaction path that is dominated by a translational movement and other surfaces will favor a path in which vibrational energy plays an important role. These different types of surfaces are often named *attractive* or *repulsive* (Fig. 25.7). In an attractive surface the saddle point occurs early and if the energy of the reactants is primarily vibrational, the reactants may vibrate apart and fall away from each other before they are able to make it over the saddle point. Alternatively if the energy of the reactants (with the same total energy as the earlier case) is translational in nature, the probability is high that the motion will carry the complex over the transition barrier to the product state. The energy remaining will often appear as increased vibrational motion in the product state. A repulsive surface has a saddle point that occurs later in the path and a reactant complex that tries to mount the barrier with a straightforward translational approach may fall back on itself because it drives straight up a high potential energy wall in a sense missing the turn onto

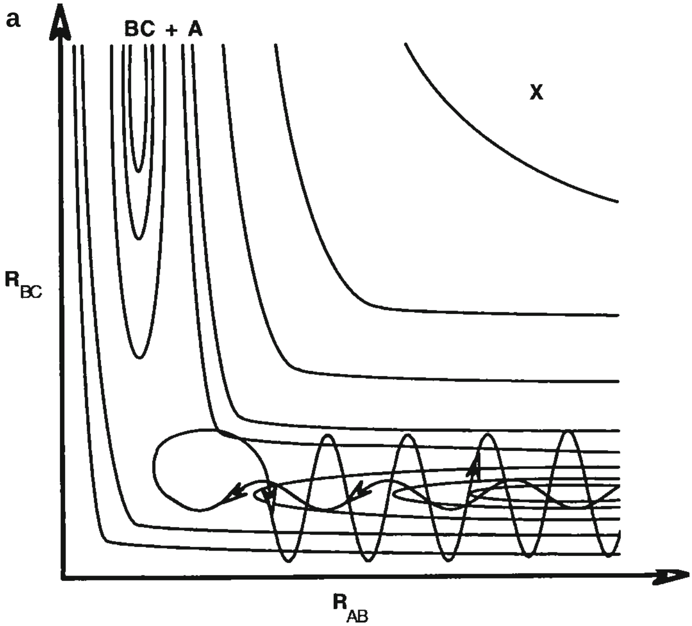
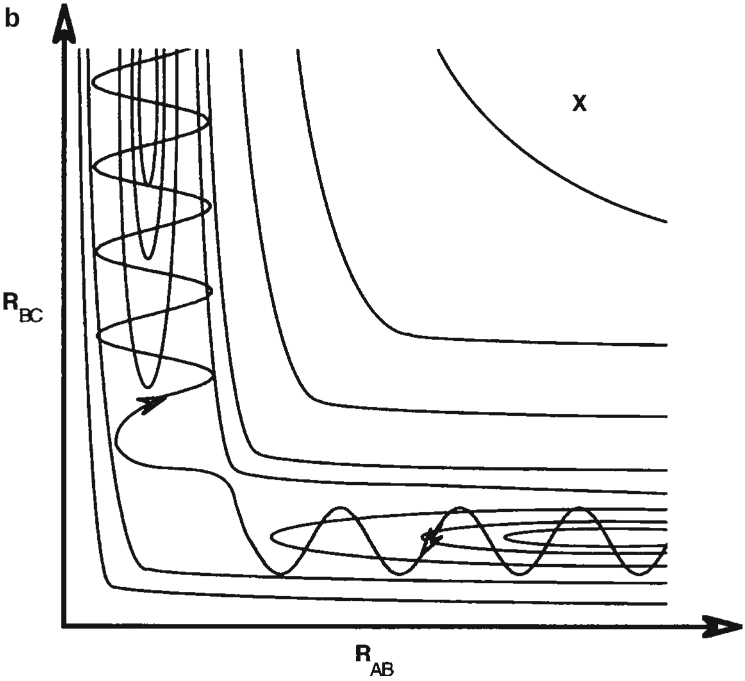


Fig. 25.6 Some of the possible trajectories for our marble-potential energy surface. [From Alberty R.A. and Silbey R.J. (1992) *Physical Chemistry*, 1st edition. Wiley, New York. Adapted by permission of John Wiley & Sons, Inc.] (**a**) This trajectory rolls off the side of the saddle point and back into the *BC* well so the reaction path is non-reactive. Note that the translational energy of *A* increases the vibrational energy of *BC* which is vibrating at a higher frequency following the encounter. (**b**) This trajectory goes over the saddle point and has a reaction probability of 1. Note how the kinetic energy in the vibratory mode changes (the frequency falls) as the transition state mounts the saddle point and then as the product is formed the kinetic energy reappears as an increased frequency of vibration. (**c**) The rate constant is related to the number of trajectories out of the total number of possible trajectories that pass through a reaction cross-section



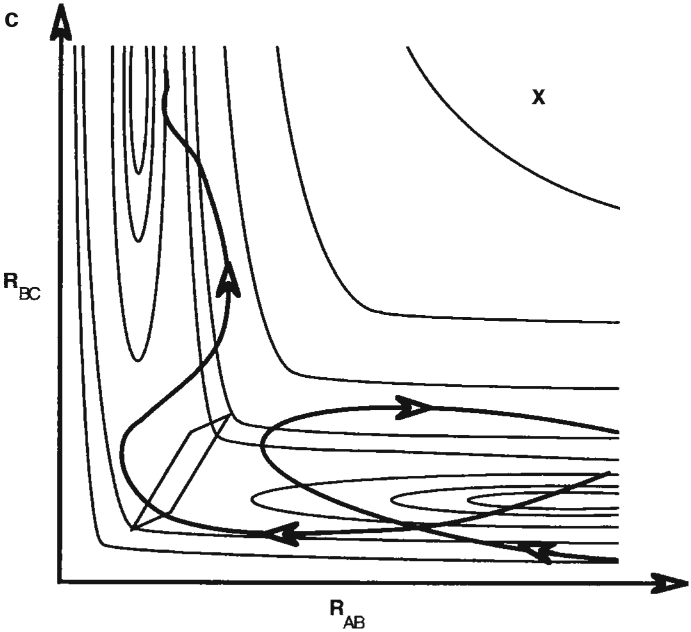


Fig. 25.6 (continued)

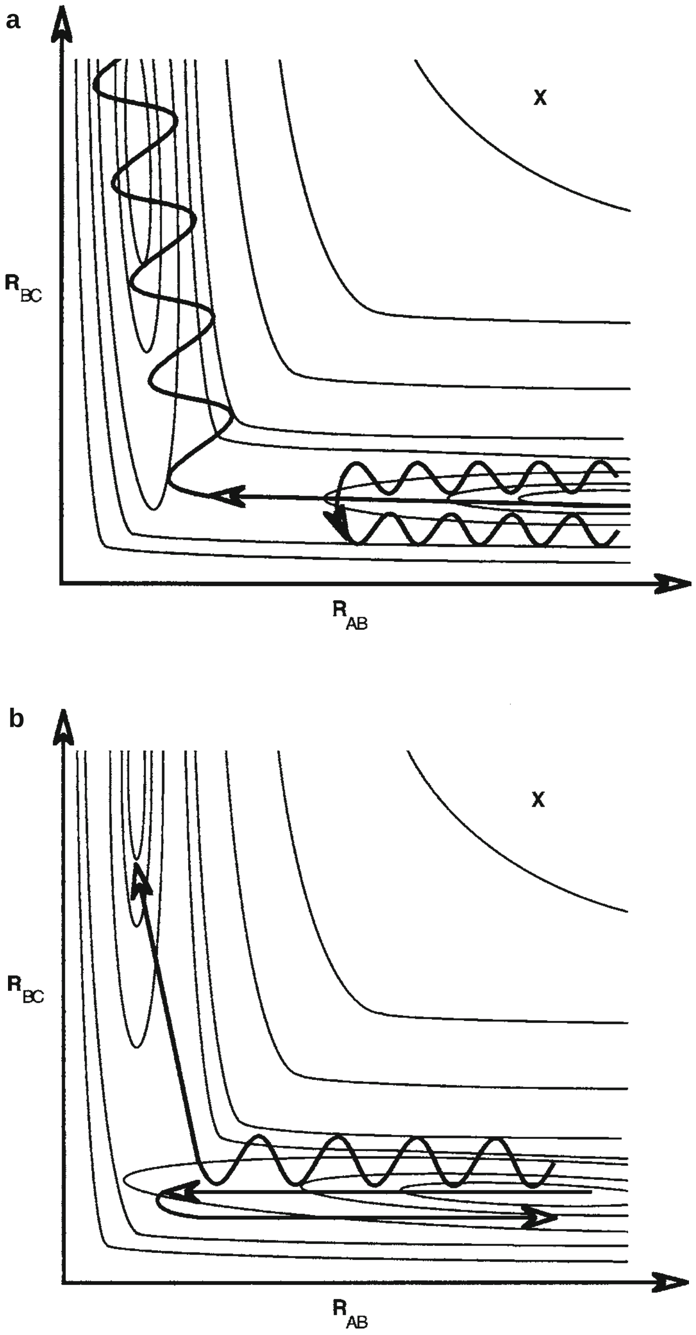


Fig. 25.7 (**a**) Attractive and (**b**) repulsive potential energy surfaces as described in the text are illustrated [From Alberty R.A. and Silbey R.J. (1992) *Physical Chemistry*, 1st edition. Wiley, New York. Adapted by permission of John Wiley & Sons, Inc.]

the minimal path and is thus unsuccessful. The reactant complex with the same amount of energy residing in vibrational modes may be successful in mounting the barrier because its motion may allow it to turn more closely into the minimal path through the saddle point. The motion of the resultant product will be translational with loss of vibratory action. It is important to recognize that the attractive and repulsive surfaces are reversed cases of each other since a reaction that is on a repulsive surface in one direction will be on an attractive surface during the return reaction.

We have used a very visual and simplified abstraction in our discussions to this point. Several points deserve mention. First of all we have been discussing only translational and vibrational modes of action but in many cases rotational motion must also be considered. This introduces no new conceptional problems but does entail another level of complexity. Second, even though the potential energy surfaces describing chemical reactions ideally would be quantum mechanical rather than Newtonian, in fact the classical approach is often *adequate* to a reasonable experimental precision. However, there are important quantum mechanical effects that must be considered in movement on the potential energy surface including tunneling or the penetration of a particle through a classically forbidden potential energy barrier.

Finally we add one more refinement to our potential energy surface picture. Reaction paths that lead over a saddle point will likely find more than one possible nearby product energy well to drop into. Thus depending on the reaction path, a variety of products may be found in varying yields for a reactant complex approaching the saddle point. The position of the potential energy wells for each of these possible product states and the trajectory with its translational and vibrational components will determine the population ratios of each of the product states. Thinking about this situation as a Chinese checker board with marbles being rolled onto it is picturesque and useful. This description explains why most chemical reactions have limited yields of the desired product; because of competing reactions. It should also be easy to see why changing the conditions of the reaction may be able to limit the trajectories. Thus the reaction preferentially proceeds toward one product well instead of another well. Knowing the potential energy surface and the conditions that lead to certain trajectories on the surface can provide a technology that can make a given reaction more efficient and economically more feasible. Recall that this was one of the reasons given earlier for the study of chemical kinetics. While a catalyst will lower the barrier potential of the transition state making the probability of a reaction more likely and thus increase the rate, it will generally accelerate the production of all the products. Given the severe competition in nature for energy and materials, the ability to be kinetically economical would likely offer an evolutionary advantage. One of the kinetic advantages of the biological catalysts, enzymes, is that they not only accelerate the rate of a reaction but also limit the number of optional product states thus achieving an economy of both rate and specificity or yield. With this as an introduction let us proceed to a discussion of the kinetic behavior of biological systems, enzymes, and surface systems.

25.8 Solution Kinetics Are More Complicated Than the Simple Kinetic Behavior of Gases

To this point we have discussed the general principles of kinetics and drew heavily on simple molecular interactions in a vacuum, in other words, gas behavior. This has allowed us to use kinetic theory to discuss the important abstraction of the potential energy surface. Yet most of the reactions of interest to the biologist and biophysical chemist occur in solution. We have drawn a fairly detailed picture of the molecular interactions that complicate the description and behavior of solutions. In general the rate of a reaction in solution cannot occur any more rapidly than the rate of diffusion which brings the reactants together. Diffusion is the result of a chemical potential driving force that affects the motion of chemical species. We have also seen that the motion of ions in solution can be affected by externally applied electric fields. Finally we have explored how the electric field at the surfaces/solvent interface leads to the important transitional structure of the interphase. We will now see how the rates of surface (binding) reactions and electrochemical processes, which are extremely common in biological systems, are very sensitive the electrical and chemical potentials at those surfaces.

25.9 Enzymes Are Macromolecular Catalysts with Enormous Efficiency

Enzymes are remarkable catalysts. Not only do they effectively accelerate the rate of a reaction they limit the potential side reactions so that the yield of a reaction is essentially 100%. The general construction of an enzyme is that of a protein with definite three-dimensional structure. This three-dimensional construction allows space for the binding of the reactant molecules in the *active site* where amino acid residues and/or other cofactors called *prosthetic* groups (Table 25.1) are carefully arranged in space (Fig. 25.8). The spatial arrangement of the active site most likely forms a potential energy surface that is complementary to the transition state of the reaction. Thus the active site serves to stabilize the transition state with the net result that the relative population of the transition state over time is increased. Not only is the ΔG of the transition state lowered however, the overall enzyme structure acts to shepherd the reactants on a narrow reaction path from reactants to specific products. While the details for different enzymatic systems depend on the specific substrates and the chemistry class of the reaction (hydrolysis, dehydrogenation, esterification, etc.), all enzymes utilize the interactional energies discussed previously to form the complementary potential energy die into which the nascent transition state is to be cast. The overall energy of the reactant interaction with the active site can be determined from the equilibrium constants of the reactant–enzyme complexes (ES complexes, vide infra). The equilibrium constants are usually on the order

Table 25.1 Prosthetic groups in enzymes

Enzyme class (enzyme)	Prosthetic group
Hemoproteins	
Hemoglobin	Fe protoporphyrin
Cytochrome *c*	Fe protoporphyrin
Catalase	Fe protoporphyrin
Flavoproteins	
Succinate dehydrogenase	Flavin adenine dinucleotide
Luciferase	Flavin adenine dinucleotide
Metalloproteins	
Alcohol dehydrogenase	Zn
Xanthine Oxidase	Mo and Fe
Superoxide dismutase	Cu and Zn (bovine)
Dopamine-β-hydroxylase	Cu
Quinoproteins	
Methylamine dehydrogenase	pyrroloquinoline quinone
Chlorophylls	Mg protoporhyrin

of 10^{-2}–10^{-8} M which corresponds to free energies of interaction on the order of –12 to –50 kJ/mol. These energies are approximately 10–15 times lower than those associated with covalent bonding.

Much of the catalytic specificity of enzymes lies in the highly precise binding of reactants or *substrates* at the active site. The complex three-dimensional structure of the enzyme polypeptide chains with the concurrent control of the local milieu is the method used to achieve such precise binding. It is easy to imagine that a transition state stabilized by an electrostatic interaction between an active site lysine and a region of negative charge density on a reactant will be energetically stabilized if water is excluded from the active site. If water is allowed in the active site the dielectric constant will rise and the coulombic interaction will be weakened. An active site where this type of interaction is needed would be expected to be in a hydrophobic peptide environment so that the likelihood of water molecules being found there is significantly decreased. Such strict rules of engagement allow for very efficient control. Many enzymes act as chemical control valves by modifying the active site thus affecting the ability of the substrate to bind precisely. Thus the catalytic power of the enzyme is modified. The modification of the complementary potential energy surface of the active site makes the active site a *control surface* in which reactions can be activated or arrested. Since the potential energy map of the control surface depends on the spatial arrangement of the residues in the active site, a modification of the spatial arrangement can be linked through the polypeptide chain to an event such as binding of product to a site on the enzyme a great distance away from the active site. Changes in the enzymatic activity because of long-distance modifications such as these are the *allosteric* interactions that we have previously mentioned. Such interactions are characteristic of enzymes and transport proteins that are regulated.

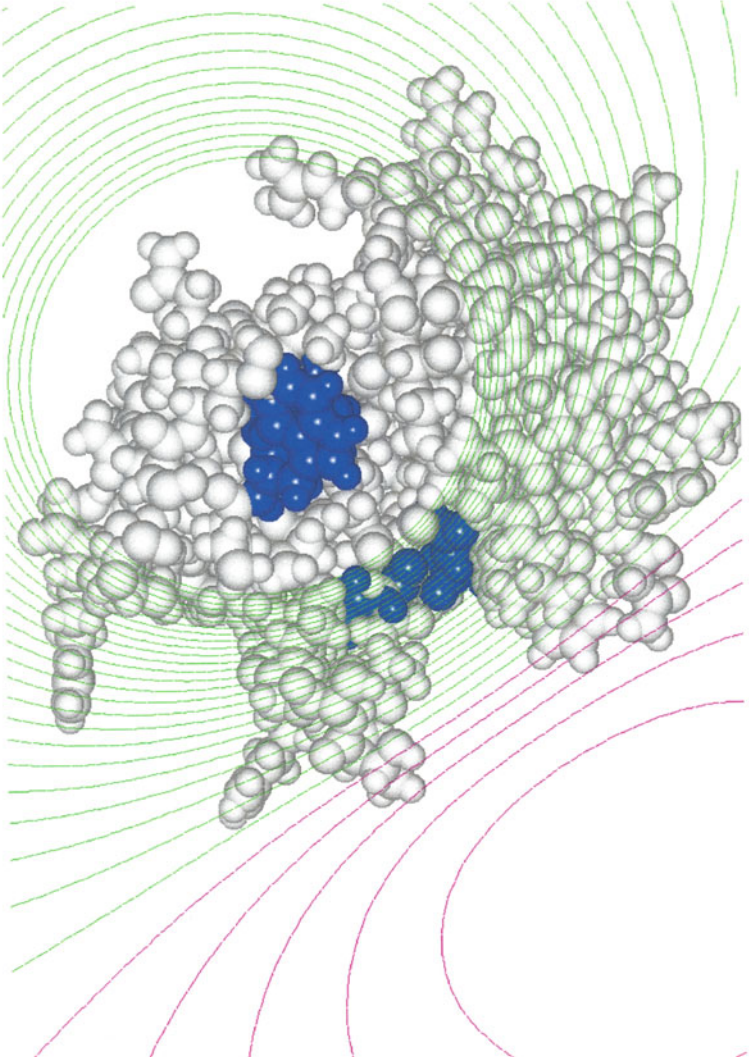


Fig. 25.8 The three-dimensional shape of an enzyme, lysozyme, taken from the x-ray crystallographic coordinates. The contour map is the electrostatic field associated with the active site. The active site residues (35, 52, 62, 63, 101, 108) were selected and then a semi-empirical quantum mechanical calculation using a CNDO technique with pseudo-fluorine parameterized classical boundaries was performed. The calculated charge distribution was then used to determine the electrostatic field which is shown as a potential energy surface that would be seen by a charge approaching the active site from 0.5 Å above the page. The electrostatic field is an important potential energy surface that will determine the likelihood of approach and orientation of an incoming ligand

25.9.1 Enzyme Kinetics

The initial step to understanding the mechanism by which an enzyme works is an analysis of its kinetics. Often kinetic study will also reveal when an enzyme is subject to allosteric control. Many enzymes are found to have an initial reaction rate **V**, that is first order with respect to substrate concentration at low [**S**]; and zero order when [S] is high (Fig. 25.9). When enzyme kinetics are studied, small fixed concentrations of enzyme (**E**) are assayed and the initial rates are used because products (**P**) are often inhibitory. In 1913, L. Michaelis and M. Menten analyzed these results and proposed a model of enzyme kinetics that accounted for the observations.

$$E + S \underset{k_1}{\overset{k_2}{\leftrightharpoons}} ES \underset{k_3}{\overset{k_4}{\leftrightharpoons}} EP \underset{k_5}{\overset{k_6}{\leftrightharpoons}} E + P \tag{25.55}$$

ES forms from *E* and *S* with a rate constant of k_1. *ES* can either fall back to *E* and *S* at rate k_2 or can proceed to *P*. Initially it is highly unlikely that any *P* will revert to *E* + *S* since little *P* as yet exists; so we ignore this possibility. This simplifies Eq. (25.55) to

$$E + S \underset{\mathrm{k}_2}{\overset{\mathrm{k}_1}{\rightleftharpoons}} ES \xrightarrow{\mathrm{k}_{\mathrm{cat}}} E + P \tag{25.56}$$

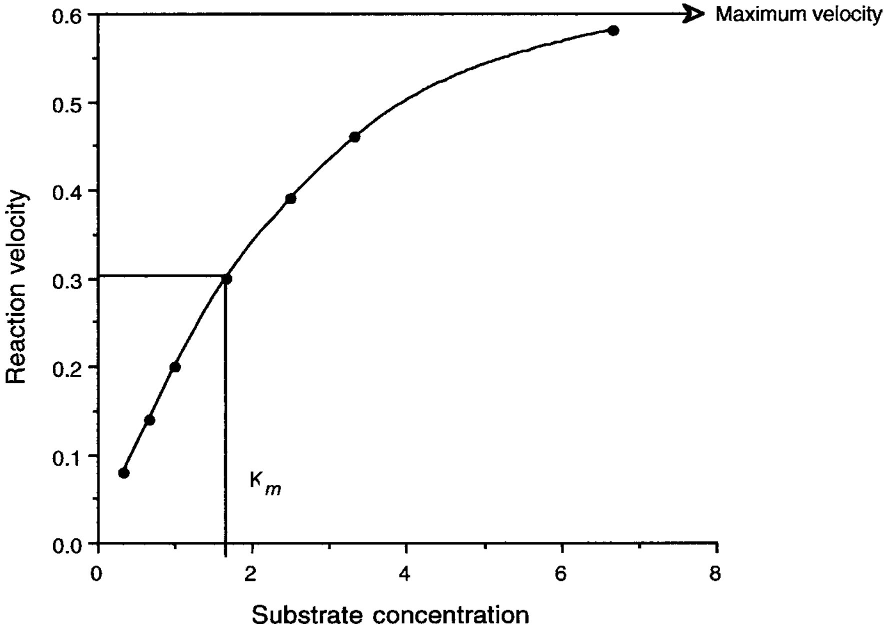


Fig. 25.9 Plot of the reaction velocity, *V*, against [*S*] for an enzyme that obeys Michaelis–Menten kinetics

The appearance of P is the product of the concentration of ES at rate constant k_{cat}:

$$V = k_{\text{cat}}\,[ES] \tag{25.57}$$

We can write $[ES]$ in terms of its formation: $k_1[E][S]$; and its breakdown: $(k_2 + k_{\text{cat}})[ES]$. If we make the steady-state approximation that the concentration of the ES complex is unchanging during the reaction, then it follows that the formation and breakdown of ES are equal:

$$k_1\,[E]\,[S] = (k_2 + k_{\text{cat}})\,[ES]\,. \tag{25.58}$$

(The steady-state approximation is reasonable because the amount of enzyme is small and is likely to be proportionately saturated at a given substrate concentration throughout the course of the experiment, therefore $[ES]$ is fixed and steady under these conditions.) From Eq. (25.58) we can find $[ES]$ by algebra

$$[ES] = \frac{[E]\,[S]}{(k_2 + k_{\text{cat}})\,/k_1} \tag{25.59}$$

The denominator composed of three constants is itself a constant for the reaction and is called the *Michaelis constant*, K_{M}:

$$K_{\text{M}} = \frac{(k_2 + k_{\text{cat}})}{k_1} \tag{25.60}$$

The $[S]$ in the numerator of Eq. (25.59) is essentially $[S]_{\text{initial}}$ because the relative concentration of enzyme is very small compared to S. $[E]$ is the total enzyme minus the enzyme complexed with substrate:

$$[E] = [E_T] - [ES]\,. \tag{25.61}$$

Combining Eqs. (25.60) and (25.61) and substituting into Eq. (25.61) gives

$$[ES] = \frac{([E_{\text{T}}] - [ES])\,[S]}{K_{\text{M}}} \tag{25.62}$$

Algebraic manipulation to solve for $[ES]$ gives the following:

$$\begin{aligned}[ES] &= [E_{\text{T}}]\,\frac{[S]\,/K_{\text{M}}}{1 + [S]\,/K_{\text{M}}}\\ &= [E_{\text{T}}]\,\frac{[S]}{[S] + K_{\text{M}}}\end{aligned} \tag{25.63}$$

We have been doing all of this to get an expression for $[ES]$ that we could substitute into Eq. (25.57) which is the catalytic rate expression. Substituting Eq. (25.63) into (25.57) gives

$$V = k_{\text{cat}}\,[E_{\text{T}}]\,\frac{[S]}{[S] + K_{\text{M}}} \tag{25.64}$$

The maximal rate of the reaction, V_{max}, occurs when the enzyme is completely saturated with substrate, i.e., every catalytic site is filled and operating. This condition occurs when $[S]$ is much higher than K_{M}. Therefore as $\dfrac{[S]}{[S] + K_{\text{M}}} \to 1$, $V \to V_{\text{max}}$, which gives

$$V_{\text{max}} = k_{\text{cat}}\,[E_{\text{T}}] \tag{25.65}$$

and

$$V = V_{\text{max}}\frac{[S]}{[S] + K_{\text{M}}} \tag{25.66}$$

This equation describes the data shown in Fig. 25.9. When the substrate concentration is low, i.e., $[S] << K_{\text{M}}$, $V = [S]\ V_{\text{max}}/K_{\text{M}}$. Under these conditions the rate is proportional to the substrate concentration. At high $[S]$ where $[S] >> K_{\text{M}}$, $V = V_{\text{max}}$. The rate is maximal and independent of $[S]$.

25.9.2 Enzymes Can Be Characterized by Kinetic Properties

The overall rate of an enzymatic catalytic process is determined by the time needed to bind substrate, undergo catalytic reaction, release product, and recycle the enzyme system in preparation for another cycle. These characteristics of an enzyme can be described by its K_{M} and V_{max} (determined as shown in Fig. 25.10). In general K_{M} reflects the binding of substrate to enzyme and V_{max} varies with the cycling or turn-over rate of the enzyme. K_{M} is equal to the substrate concentration at which $V = \dfrac{V_{\text{max}}}{2}$. This occurs when $[S] = K$. K_{M} thus reflects the activity of substrate needed to fill half the active sites and is dependent on substrate and physical conditions such as pH, temperature, and ionic strength. Given the K_{M}, the fraction of active sites occupied can be calculated for any $[S]$:

$$f_{\text{ES}} = \frac{V}{V_{\text{max}}} = \frac{[S]}{[S] + K_{\text{M}}} \tag{25.67}$$

As Eq. (25.60) shows, K_{M} is a measure of the movement away from ES $(k_2 + k_{\text{cat}})$ divided by the movement toward ES (k_1). The numerator can be dominated by the dissociation into reactants rather than products, $k_2 >> k_{\text{cat}}$, which makes $K_{\text{M}} = \dfrac{K_2}{K_1}$. This expression is the equilibrium constant for $E + S \to ES$ in terms of the rate constants:

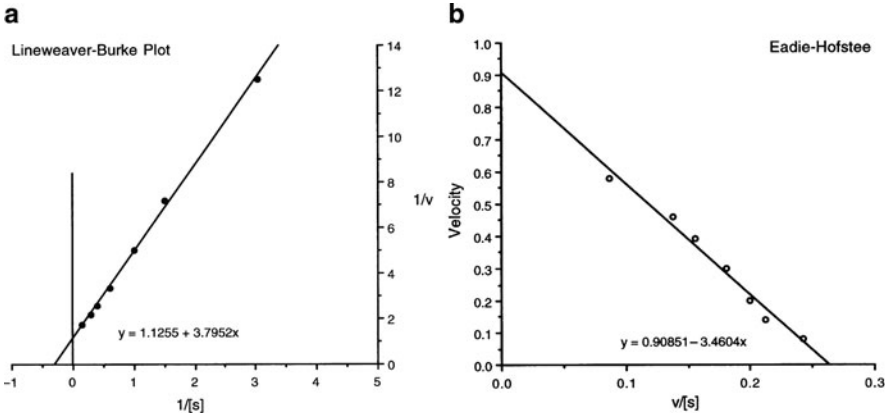


Fig. 25.10 Plots used to determine the kinetic parameters in enzyme studies. (**a**) The double reciprocal plot (Lineweaver–Burk) yields a straight line. The x intercept gives $-\frac{1}{K_m}$ and the y intercept $= \frac{1}{V_{\max}}$. (**b**) The Eadie–Hofstree plot graphs v against $\frac{v}{[S]}$ which yields a straight line with the x intercept of $\frac{V_{\max}}{K_m}$ and the y intercept at $V_{\max}$

$$K_{\mathrm{M}} = \frac{k_2}{k_1} = \frac{[E]\,[S]}{[ES]} = K_{ES} \tag{25.68}$$

For a potential energy surface where the dissociation to reactants is preferred over proceeding on to products, K_{M} represents the strength of the ES complex with a large K_{M} indicating weak binding of reactants with the enzyme and a small K_{M} reflecting a strong binding. The K_{M} reflects the affinity of the enzyme for substrate and thus reflects an important parameter of enzyme activity, the enzyme's ability to be active at a certain substrate concentration. If the K_{M} is very small then the enzyme will catalyze reactions at very low concentrations and may thus appear to be skirting the law of mass action. When the substrate concentration is below the K_{M} the effects of the law of mass action will be seen. In the early study of enzymes the apparent exception to thermodynamic laws such as mass action and the high specificity of the reaction path lead to questions regarding the possible exception of enzyme driven reactions to the second law of thermodynamics. Careful thermodynamic experiments clearly demonstrated that enzymes are to complete compliance with all of the thermodynamic laws.

While an enzyme's tendency to bind substrate and begin its catalytic work is measured by the K_{M}, the speed of the actual catalytic process that produces and releases product is reflected in its *turn-over number*. The turn-over number is equal to k_{cat}. It is the number of substrate molecules converted to product in a unit time when the enzyme is fully saturated. In other words it is equal to $V_{\max}$ divided by a proportionality constant. This constant is the concentration of active sites, $\mathbf{[E_T]}$:

$$V_{\max} = k_{\mathrm{cat}}\,[E_T] \tag{25.69}$$

The enzyme beef-lever catalase which promotes the reaction

$$H_2O_2 \rightarrow O_2 + 2H_2O \tag{25.70}$$

has a turn-over number of 4.67×10^4 which means that each catalytic cycle takes 21.4×10^{-6} s!

Given the condition of substrate saturation the rate of catalysis is equal to k_{cat}. This requires that $[S]$ is much greater than K_{M}, which is a condition that is seldom met physiologically. Typically the $\frac{[S]}{K_{\mathrm{M}}}$ ratio ranges from 0.01 to 1 and the great majority of active sites remain unoccupied. If $[S] << K_{\mathrm{M}}$ the rate will be much less than k_{cat}. because of the low saturation of active sites. Under these conditions the rate can be determined from

$$V = \frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}[E][S] \tag{25.71}$$

Under the conditions of $[S] << K_{\mathrm{M}}$ the $[E]$ is essentially that of the total enzyme, $[E_{\mathrm{T}}]$ and so the velocity depends on the ratio $\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}$ and $[S]$. This ratio $\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}$ is a measure of the kinetic efficiency of an enzyme. Remember that K_{M} contains the rate constants for the formation and dissociation of the ES complex. If k_{cat} is much greater than k_2, the rate of dissociation of the ES complex, then the velocity and the value of $\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}$will be limited by k_1, the rate of formation of the ES complex. k_1 is also physically limited since the ES complex cannot form any faster than the substrate and enzyme can be transported and encounter one another. In a field-free phase, this rate of transport is set by diffusion. The diffusion limited value of k_1 is between 10^8 and 10^9 $\mathrm{M}^{-1}\mathrm{s}^{-1}$. The $\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}$for catalase is 4×10^7 $\mathrm{M}^{-1}\mathrm{s}^{-1}$ so it is an extremely efficient enzyme operating almost at the diffusion controlled limit. The inquisitive reader should be flipping to Table 25.2 and grabbing a calculator right now because an enzyme may reach a diffusion limited state while having a K_{M} or k_{cat} that are markedly different. For example comparison between carbonic anhydrase and GTP-cyclohydrolase show that both have a very high $\frac{k_{\mathrm{cat}}}{K_{\mathrm{M}}}$of 7.5×10^7 and 5×10^8 $\mathrm{M}^{-1}\mathrm{s}^{-1}$, respectively, indicating that their catalytic efficiency is very high and it is a physical step, that of diffusion, not a chemical step which is rate limiting. This means that there is nothing the enzyme can do to increase its catalytic power. Alberty and Knowles have argued that such enzymes have reached the end of their evolutionary development as catalysts. However, if an enzyme is placed in a system in which it is no longer constrained by diffusion, the rate can be increased still further. This is the approach taken in the construction of enzyme complexes which have achieved factors 1000 times higher rates by careful arrangement of enzymes such that an enzyme is able to act on the product of a previous enzyme as it is channeled almost like a fast assembly line. In electron transport systems the

Table 25.2 K_M and k_{cat} values for selected enzymes

Enzyme (substrate)	K_M(M)	Turn-over #/s
Catalase	1.2×10^{-3}	46,667
Carbonic anhydrase		
CO_2	8×10^{-3}	600,000
GTP-cyclohydrolase		
GTP	2×10^{-8}	10
Chymotrypsin		
Acetyl-l-tryptophanamide	5×10^{-3}	100
Lysozyme		
Hexa-*N*-acetylglucosamine	6×10^{-6}	0.5
Penicillinase		
Benzylpenicillin	5×10^{-5}	2000
Pyruvate decarboxylase		
HCO_{3-}	1×10^{-3}	
ATP	6×10^{-5}	
Pyruvate	4×10^{-4}	
Triose phsophate isomerase		
Glyceraldehyde-3-phosphate	1×10^{-5}	1000

redox centers are closely arranged to supersede the diffusion barriers. These systems also use structural proximity to take advantage of the ability of electrons to tunnel through classical barriers which increases the kinetic rate to an even higher level. The rate of charge transfer in these transport chains is on the order of 10^{12} s^{-1}. Another point can be made from this example of catalytic efficiency. When a substrate is held loosely, the turn-over rate can be quite high and enzymes such as these functions efficiently in high concentrations of substrate but will be inefficient in low [S] environments. An enzyme with a high affinity for substrate must pay with a lower rate of turnover but is a highly efficient catalyst in very low [S] environments.

25.9.3 Enzymes Are Complex Systems Subject to Biophysical Control

We have discussed how alteration in the potential energy surface of the active site can act to control the energy of the transition state of the *ES* complex. Another control mechanism in enzymes is the physical gating that would be reflected in k_1 and is another physical barrier to catalysis in addition to diffusion. Other portions of the three-dimensional protein structure can act as gates or selection filters for substrates and thus control the overall rate of the enzyme. Control of enzymatic activity is quite important in a physiological system and can be applied

(1) at the active site,
(2) at a regulatory site that alters the control surface of the active site, or
(3) at the access regions to the active site that allow the substrates to arrive or the products to leave the active site region (essentially a toll gate or a bottleneck control system).

Inhibition of enzyme activity can be either *reversible* or *irreversible*. Irreversible inhibitors usually bind very tightly and dissociate from the enzyme very slowly if at all. Reversible inhibitors characteristically dissociate very rapidly from the enzyme complex thus having a very limited period of inhibitory effect.

Reversible inhibition of Michaelis–Menten type enzymes comes in three major types. Each is distinguishable by experiment and has a characteristic pattern on double reciprocal plots (Fig. 25.11). In *competitive* inhibition a molecule mimics the transition state. It will fit into the active site and occupy a catalytic site but is itself unreactive. Normal substrate is thus blocked from being bound and processed into product. These inhibitors can be reversed by diluting out the inhibitor or increasing the amount of true substrate available. Essentially a new competing step

$$E + I \rightleftharpoons EI \tag{25.72}$$

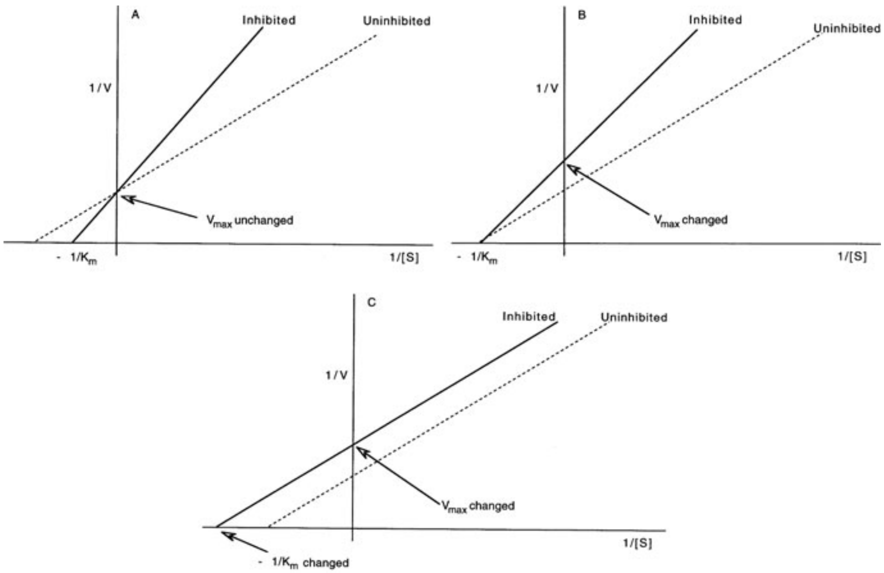


Fig. 25.11 Plots of reversible inhibition patterns for Michaelis–Menten type enzymes. (**a**) competitive; (**b**) non-competitive; (**c**) mixed or uncompetitive

is added that operates in parallel to the reaction of Eq. (25.56)

$$\begin{array}{l} E + S \rightleftharpoons ES \rightleftharpoons E + P \\ \ + \\ I \\ \downarrow \\ EI \end{array} \tag{25.73}$$

In the presence of a reversible transition-state analog, there is no change in $V_{\max}$ for the enzyme. This is because the active sites can be saturated with natural reactants in the presence of a high enough concentration of substrate. However, the K_{M} will be increased reflecting the need for a higher substrate concentration to reach $\frac{1}{2}V_{\max}$. The inhibitor essentially decreases the catalytic efficiency by shifting K_{M} and not k_{cat}. The new K_{M} can be found by writing the total enzyme concentration and K_{I}, the inhibitor dissociation constant, as

$$[E_{\mathrm{o}}] = [E] + [ES] = \text{and } K_I = \frac{[E]\,[I]}{[EI]} \tag{25.74}$$

$$K_{\mathrm{M}'} = K_{\mathrm{M}}\left(1 + \frac{[I]}{K_I}\right) \tag{25.75}$$

The experimental demonstration that compounds that look like the transition state are competitive inhibitors is one of the lines of evidence that supports L. Pauling’s suggestion in 1946 that the active site was structured to be a complementary transition state and hence achieved its catalytic success. (Another line of evidence supporting this assertion is the demonstration of catalytic activity by antibodies raised against a transition state itself, the antibody then acting as an active site configuration). A final point can be made linking the transition state and the structural relationship between the reactant and product. Multiple examples of the competitive inhibition of an enzyme by its product clearly show the structural lineage from transition state to product. This is a phenomenon that nature has taken advantage of by using excess product production as a control to attenuate an enzymes activity.

Not all inhibitors act at the active site. Certain inhibitors bind elsewhere on the enzyme and interfere with the catalytic action but not with the binding of substrate. Inhibition of this sort will not affect the K_{M} but will decrease the $V_{\max}$ of the enzyme in the presence of inhibitor regardless of substrate concentration. Inhibitors of this category are named *non-competitive*. Finally inhibitors that effect catalytic efficiency by changing substrate binding as well as affecting $V_{\max}$ can be kinetically distinguished and are called *uncompetitive*.

Further Reading

General

Good treatments of general kinetics and enzyme kinetics are the logical places to explore beyond our qualitative treatment are to be found in the chapters in the textbooks listed in the second chapter of this book.

Further Aspects of Enzyme Kinetics

Klinman J.P. (1978) Kinetic isotope effects in enzymology. *Adv. Enzymol.*, **46**:415–494. (This important effect in the study of kinetics requires an appreciation of the quantum mechanics operating at the active site of enzymes.)

Lerner R.A. and Tramontano A. (1988) Catalytic antibodies, *Sci. Am.*, **258, 3**:58–70. (Evidence that making a chemical “mold” of the transition state can act as a catalytic site itself.)

Natarajan K.R. (1991) Biocatalysis in organic solvents, *J. Chem. Educ.*, **68**:13–16. (What is the water environment for the catalytic activity of an enzyme?)

Slater E.C. (1981) Maxwell demons and enzymes, *TIBS,* Oct:280–281. (The enzyme a perpetual motion machine?)

Walsh C. (1979) *Enzymatic Reaction Mechanisms*, W.H. Freeman, New York. (This wonderful book is filled with descriptions of a wide variety of enzyme mechanisms. It is like a chemical gothic mystery filled with fascinating characters.)

Interesting Directions to Go

Bluestone S. and Yan K.Y. (1995) A method to find the rate constants in chemical kinetics of a complex reaction, *J. Chem. Educ.*, **72**:884–886.

Greene E.F. and Kuppermann A. (1968) Chemical reaction cross sections and rate constants, *J. Chem. Educ.*, **45**:361–369. (This article is an excellent introduction to the ideas of reaction cross sections in chemical kinetics)

Loudon G.M. (1991) Mechanistic interpretation of pH-rate profiles, *J. Chem. Educ.*, **68**: 973–984. (This article discusses the kinetics of acid-base reactions and its use in determining mechanisms. The examples given are biochemical reactions.)

Tagg S.L., LeMaster C.L., LeMaster C.B., and McQuarrie D.A. (1994) Study of the chaotic behavior of a damped and driven oscillator in an asymmetric double-well potential, *J. Chem. Educ.*, **71**:363–274. (Application of chaos modeling to a chemical system. A good place to start learning the vocabulary as it applies to chemical kinetics.)

Truhlar D.G. and Gordon M.S. (1990) From force fields to dynamics; classical and quantal paths, *Science*, **249**:491–498.

Problem Sets

1. What is the effect on the rate of a reaction with an activation energy of 25 kJ/mol when the temperature increases 1°? 5°? 10°? 20°? 50°? 100°?
2. Calculate the change in rate constants for the following activation energies with the accompanying temperature changes:

(a) 50 kJ/mol from 0 to 10°C
(b) 5 kJ/mol from 0 to 10°C
(c) 500 kJ/mol from 0 to 10°C

3. A reaction has the following rate constants:

$$\begin{aligned} \mathrm{A} \rightarrow \mathrm{B} \quad K_1 &= 10^{-5}\mathrm{s}^{-1} \\ K_{-1} &= 10^{-8}\mathrm{s}^{-1} \end{aligned}$$

(a) What is the equilibrium constant for the reaction?
(b) Which way will the reaction go if $[A] = 2.6 \times 10^{-6}$ M and $[B] = 1.5 \times 10^{-4}$ M?
(c) The reaction will proceed in which direction if $[A] = 1.5 \times 10^{-5}$ and $[B] = 1.5 \times 10^{-1}$ M?
(d) How long will it take to reach equilibrium for this reaction?

4. You have received a shipment of nucleotide labeled with the β emitter ^{32}P with a specific activity of 10 mCi/mol (1 μCi = 2.22×10^6 disintegrations/min). The specific activity was measured by the isotope manufacturer 32 days ago. You will use 1 mmol of radionuclide in your experiment. You expect to be able to label each molecule in your reaction with five atoms. It will take 5 days to complete the assays before measuring can begin. Your detection system will require 4×10^7 decays to be able to record any data. The half-life of ^{32}P is 14.2 days.

(a) How many moles of product will you need in order to perform your experiment?
(b) If you bought fresh radioisotope of the original specific activity, how much product would be required to perform your assay?

5. The half-life of ^{14}C is 5568 years. This isotope is produced tin the atmosphere by the interaction between CO_2 and cosmic rays. There is a steady-state concentration of ^{14}C in the air. Plants that use the $^{14}CO_2$ incorporate the isotope into their molecular structure with no further additions or subtractions except by radioactive disintegration. If a sample of wood from a tree is found to have the ^{14}C content of 86% compared to atmospheric ^{14}C (assume that this number is constant), what is the age of the tree?

6. The following data is collected for an enzyme:

Substrate concentration (mmols)	Initial velocity (μmol/s)
1.716	0.416
3.432	0.728
5.20	1.040
8.632	1.560
13.00	2.028
17.316	2.392
25.632	3.01

a) What is the K_m for this enzyme?
b) What is V_{max}?

7. The binding of a substrate to an enzyme increases its reaction velocity by a factor of 1000.

   a) What is the change in the free energy of activation?
   b) Another enzyme accelerates the same reaction by a factor of 10,000. What is the energy of activation of this enzyme−substrate complex?

8. From Table 25.2 determine the $\frac{k_{cat}}{\mathrm{K_M}}$ for the following enzymes. What is the likely rate-limiting step in the catalytic process.
   Chymotrypsin; Carbonic anhydrase; Catalase; Penicillinase; Lysozyme
9. There are four dominant methods of chemical catalysis that occur in biological systems.

   (1) Entropic or proximity catalysis
   (2) Electrophilic and nucleophilic attack
   (3) General acid–base catalysis
   (4) Strain and distortion
   Name an example of each and discuss in terms of the potential energy surface the forces involved in the catalytic mechanism.

Chapter 26
Dynamic Bioelectrochemistry – Charge Transfer in Biological Systems

Contents

26.1 Electrokinetics and Electron Charge Transfer Depend on Electrical Current Flow in Biochemical Systems

If electrostatic forces define the chemical shape of our world, to a great degree it is the movement of charge that brings that world to life. The movement of charge is necessary for sensation, mental activity, muscular movement, and energy transduction by photosynthesis and metabolism to name just several examples. When

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_26, © Springer Science+Business Media, LLC 2010

an electric field exists either in free space or in a conducting medium, charges will move under the field’s influence. This movement of charge is the *electric current*. The principles of conduction and ionic current have been introduced already in Chapter 23 and Appendix P reviews the fundamentals of electrical circuits. Now we will explore the movement of charge in biochemical systems of multiple phases (electrokinetics) and in a matrix of biochemicals (electron transfer).

26.2 Electrokinetic Phenomena Occur When the Elements of the Biological Electrical Double Layer Experience Either Mechanical or Electrical Transport

Although the cell can be thought of as a pair of phases, one intracellular and one extracellular, separated by a membrane, the membrane itself can be considered a phase. When the membrane is conceived as a phase, the cell is described by an extracellular aqueous phase in contact with a lipid phase that contacts another aqueous phase. However, the intracellular environment should probably not be considered as a unitary aqueous phase except in the most unusual of circumstances. The structure of virtually all cells, except the mature mammalian red cell, is comprised of multiple membranes and intracellular organelles packed densely together. Each membrane has associated with it two interphase regions, one on each side, representing its association with the phases. The nature of the interphase region that is associated with biological membranes is an extremely important aspect of the surface chemistry in biological systems and is also exceedingly complex.

The exact structure and role of the interphase in biological systems are still under investigation, but it surely plays an important role in the transfer of charge in the electron transport chain, the behavior of cell–substrate interaction, the adsorption of hormones and other chemical messengers on the cell surface, and the process of generating, maintaining, and altering the membrane potentials in both excitable and non-excitable cells, to name just a few examples. The structure of the double-layer region gives rise to a number of important effects associated with motion relative to the double layer, called *electrokinetic effects*.

26.2.1 The Zeta Potential Is Measured at the Outer Helmholtz Plane of the Electrical Double Layer

Charge located in the compact layer is bound tightly and, like the irrotational water, is essentially part of the same kinetic entity as the interface. In other words, it will move in lockstep with the movement of the interface. On the other hand, charge found in the diffuse layer is not held as tightly to the surface. In a sense, the ions making up the countercharge in the Gouy–Chapman diffuse layer have split loyalties to the interphase region and the bulk solution. When the solution phase and interphase are at rest (relative to one another), the ions making up the diffuse layer are held by the forces from the electrified interface (moderated of course by thermal randomizing forces). However, if either the solution or the surface move relative to the other, the ions located in the diffuse layer can be torn away from the interphase region and move with the bulk phase. What are the consequences of removing this charge from the interphase region? First of all, a portion of the charge that was needed to neutralize the field associated with the electrified interface is now gone. If a measurement were made across the complete interphase region, the interphase would no longer be electrically neutral. Furthermore, the charge that was removed from the diffuse region is now associated with the bulk phase; it is also no longer electroneutral. When charge is separated, a potential difference is the result. The motion of the bulk phase with respect to the surface therefore has resulted in a new potential field, the *electrokinetic potential.*

The electrokinetic potential is named the *zeta potential* and given the symbol ζ. The magnitude of ζ is found by measuring the potential difference between the portion of the diffuse layer that can be removed by a shearing force and the bulk of the solution (which represents the reference zero or reference ground). It is generally considered possible to remove the entire diffuse layer under a kinetic shearing force, thus the shear plane will be located just outside the outer Helmholtz plane. The zeta potential therefore almost always has the same value as ψ at the OHP. Knowledge of ψ can be used to indirectly infer some information about the surface charge excess and the compact layer. However, caution must be used, since the potential at the OHP is by no means specific for a particular interface structure or even for the charge on the surface.

There are four electrokinetic phenomena. Two of these occur when an electric field is applied to the system and cause movement of either the particle or the solvent associated with the double layer, these are *electrophoresis* and *electro-osmosis,* respectively. The other two phenomena, the *streaming potential* and the *sedimentation potential,* are instances of the generation of an electric field when the solvent or the particle, respectively, generates a potential secondary to the shearing forces during its movement. Each of these phenomena will be summarized briefly below.

26.2.2 A Streaming Potential Results When Fluid Flows in a Cylinder

When a pressure differential is applied to an electrolyte (or any fluid) in a pipe or capillary, the fluid will move through the opening with laminar flow. As the electrolyte begins to flow, its motion is opposed by the viscous force of the solution. The velocity of flow will increase until the force exerted through the pressure gradient is opposed equally by the viscous force. The velocity of flow for any given lamina increases with the distance away from the wall of the vessel. The lamina of fluid nearest the surface of the capillary will not move and will remain fixed to the capillary wall. This is consistent with the interpretation that this layer is represented by the solution phase inside the immobile outer Helmholtz plane. The first lamina of solution that is free to slip past the immobile lamina includes a portion of the diffuse Gouy–Chapman layer. The movement of the volume elements that contain Gouy–Chapman charge leads to the condition where the flowing lamina of electrolyte carries a charge. A separation and accumulation of charge occurs, and as a result, an electrical potential is generated. This voltage is the *streaming potential*. If the amount of charge drawn away from the double layer is measured as it moves through a plane normal to the surface, then a streaming current can be found. The presence of an electric field in the electrolyte causes a current to flow in the opposite direction of the streaming current. At steady state, the streaming current and the conduction current that opposes it are equal but opposite in sign. The streaming potential, E_s, is found to depend on the viscosity (η) and conductivity (κ) of the electrolyte, the pressure gradient (P) in the capillary, and the zeta potential (ζ), which is usually considered to be synonymous with ψ at the OHP. The relationship can be written as

$$E_\mathrm{s} = \frac{\zeta\varepsilon_o\varepsilon P}{\eta\kappa} \tag{26.1}$$

The streaming potential will be expressed in volts if the zeta potential is also expressed in volts.

26.2.3 Electro-osmosis Is the Transport of Solvent Coincident with Electrical Induced Flux of Electrolytes

Electro-osmosis is essentially the paired opposite of the streaming potential phenomenon. A similar system to that described for the streaming potential is considered. If an electric field is applied parallel to the surface, there will be an induction of charge transport through the electrolyte solution. The charge will be carried principally by the ions in the solution, including some of the ions that are involved in providing the diffuse Gouy–Chapman layer at the surface. As the ions are stripped from the diffuse layer and experience net flux in the parallel electric field, they carry water with them. The movement of water and ions results in a net flux of solvent whose flow is countered by a viscous force identical to that experienced when the driving force is a pressure differential. Consequently, the electrical potential gradient behaves like a gradient of pressure in a capillary or pipe, causing net flow. The mathematical relationship for a pressure differential causing current to flow (the streaming potential) and that for a current causing a net flux of fluid (electro-osmosis) can be shown to be equivalent, thus demonstrating the reciprocity of these two processes. The reciprocal equivalence of these phenomena is important because it fulfills the Onsager reciprocity relationship for irreversible processes. The relationship of the electric field to the volume of fluid flow (V) is described by

$$V = \frac{\zeta\varepsilon_o\varepsilon I}{\eta\kappa} \tag{26.2}$$

where *V* is the volume transported in unit time and *I* is the current flowing in the test system. The electro-osmotic mobility, μ_{E}, is defined as the velocity of flow of the solvent with respect to the external field and is given as

$$\mu_{\mathrm{E}} = \frac{v}{E} = -\left(4\pi\varepsilon_o\right)\frac{\varepsilon\zeta}{4\pi\eta} = -\frac{\varepsilon_o\varepsilon\zeta}{\eta} \tag{26.3}$$

The negative sign indicates that the flow of solvent is toward the electrode of the same polarity as the zeta potential. This occurs because the charge in the diffuse layer is opposite in sign to the potential at the OHP. Electro-osmotic processes are important in the behavior of ion-exchange membranes and may well play important roles in behavior across biological membranes as well. Electro-osmosis is an important cause of artifact in the electrophoretic method of capillary electrophoresis.

26.2.4 Electrophoresis Describes the Motion of Particles with an Electrical Double Layer in an Electrical Field

For the biochemist, the most familiar phenomenon associated with the electrified interface is the process of *electrophoresis*. In electrophoresis, a particle must be large enough to have a double layer; most macromolecules, organelles, and cells fit this requirement. Electrophoresis is like electro-osmosis in that an external electric field causes relative motion of the surface and the diffuse layer; it differs in that the solvent is now stationary and the particles move. Consider a system of small particles (just how small will be considered shortly) in solution that are subjected to an electric field. Streaming of the liquid secondary to electro-osmotic effects is prevented. The particles are found to move in the field. Shortly after the application of the electric field, a steady-state velocity of the particles is established with the particles moving toward one of the electrodes. The particles move because of the electrostatic interaction between the field and their own charge. They reach steady-state velocity when the electrostatic forces are opposed equally by the viscous forces associated with their movement. The velocity of the particles with respect to field strength is the electrophoretic mobility, μ, and has units of $\mathrm{m^2\ s^{-1}\ V^{-1}}$:

$$\mu = \frac{v}{E} \tag{26.4}$$

Electrophoresis refers to particles that are large enough to have an electrified interface associated with them. Most quantitative analyses are derived based on the condition that the particle is large compared to the dimension of the diffuse double layer as given by κ^{-1}, which makes the conditions similar to those for the derivation of the electro-osmotic effect. The particle and the compact layer of the electrified interface move in concert as a kinetic entity, and consequently it is once again the potential at the OHP or ζ that determines the charge associated with the particle. This leads to several interesting effects. First, since the diffuse layer has a sign opposite to that of the particle it will be attracted to the opposite electrode. Since the diffuse layer of charge wants to stay associated with the particle (which is being accelerated in one direction), yet is itself being accelerated in the opposite direction, the ion cloud opposes the motion of the particle. Ultimately, it is the particle that experiences net motion, but at the cost of decreased mobility. Like ions in an electric field, the particle does not actually drag its diffuse layer with it but leaves a portion behind and continually rebuilds the diffuse layer in front as it moves. Second, although most particles of biological interest (e.g., cells, proteins) have a net charge, a particle with no net charge can also experience electrophoretic effects, because the electrophoretic mobility is determined by the charge at the OHP, and the specific adsorption of ions at the IHP (usually anions) can lead to a net charge at the OHP. The negative mobilities of colloidal gold and oil droplets as shown in Table 26.1 are examples of this effect.

Table 26.1 Electrophoretic mobilities for selected colloids at pH 7.0

Colloidal species	Mobility ($m^2/s/V \times 10^{-8}$)
Human erythrocytes	−1.08
Human platelets	−0.85
Colloidal gold	−3.2
Oil droplets	−3.1

The derivation of a simplified expression for electrophoretic behavior can provide a glimpse of some of the complexities in a biological system. The starting point is highly simplified. A system is imagined where a spherical charged particle can have its electrified interface switched on and off and the solvent can be switched from a non-conducting solution to a conducting electrolytic one at will. Starting initially in the non-conducting solvent without an electrified interface, the particle is subjected to an electric field, *E*. The force experienced by the particle is given by

$$F = z_i e_o E \tag{26.5}$$

The velocity of the particle, *v*, depends directly on the electrostatic force and inversely on the frictional coefficient, *f* of the particle. Therefore, the following can be written as

$$v = \frac{z_i e_o E}{f} \tag{26.6}$$

Earlier, the electrophoretic mobility, μ, was defined as *v*/*E*; this can be combined with Eq. (26.6) to give

$$\mu = \frac{z_i e_o}{f} \tag{26.7}$$

Since the particle is spherical, the frictional coefficient may be calculated from Stokes' law, and Eq. (26.7) may be rewritten as

$$\mu = \frac{z_i e_o}{6\pi \eta r} \tag{26.8}$$

where r is the radius of the particle. This is a general equation for the simplified system described. In the more realistic case, it will be found that two approaches to the same problem give two answers. Understanding this apparent discrepancy leads to some important insights.

The switch is thrown and the solvent becomes a conducting electrolyte and the electrified interface is allowed to form. What will have changed? First of all, an electrified double layer exists that contributes its own field to that of the external field. Use of ζ rather than the charge $z_i e_o$ on the particle will yield more accurate results. Second, the size of the diffuse layer and the contribution it makes to the apparent viscous drag on the particle will vary. As the ratio of the reciprocal length to the radius of the particle changes, the importance of the drag contribution by the diffuse layer will vary. It has been indicated that electro-osmosis and electrophoresis share fundamental similarities. One of the most important is the derivation of the electro-osmotic and electrophoretic mobilities. It can be shown that if the radius of the particle is large relative to the reciprocal length of the diffuse layer, then the electrophoretic and electro-osmotic mobilities differ only in sign. The Smoluchowski equation for electrophoretic mobility gives the result:

$$\mu = (4\pi \varepsilon_o) \frac{\varepsilon \zeta}{4\pi \eta} = \frac{\varepsilon \varepsilon_o \zeta}{\eta} \tag{26.9}$$

In 1924, Hückel derived an expression for the electrophoretic mobility which gave the result:

$$\mu = \frac{2\varepsilon_o \varepsilon \zeta}{3\eta} \tag{26.10}$$

These are two significantly different answers for the same problem. Interestingly, it turns out that both are correct for a specific set of conditions. How can this apparent discrepancy help in the understanding of the processes in electrophoresis?

It is entirely feasible to derive an equation that sets the electric force on the particle in this system equal to the viscous force when the particle is at steady-state conditions. In like fashion to Eq. (30.8), the viscous force, f_v, can be found through Stokes' law:

$$f_\mathrm{v} = 6\pi v a \eta \tag{26.11}$$

where v is the velocity of the particle and a is the particle radius. The electrical force, f_e, is given simply as

$$f_{\mathrm{e}} = QE \tag{26.12}$$

Q is the charge and E is the electric field. Combining these equations in a manner analogous to that described for Eqs. (26.6) through (26.8) gives the following result for μ:

$$\mu = \frac{Q}{6\pi a\eta} \tag{26.13}$$

Q is evaluated to take into account the charge on the double layer and the radius of the particle, including the compact layer, and is written as

$$Q = (4\pi\varepsilon_o)\,\varepsilon a\,(1 + \kappa a)\,\zeta \tag{26.14}$$

Combining Eqs. (26.13) and (26.14) gives the following result:

$$\mu = \frac{(4\pi\varepsilon_o)\,\varepsilon\zeta}{6\pi\eta}\,(1 + \kappa a) \tag{26.15}$$

The term κa relates the radius of the particle and the thickness of the diffuse layer. In the limit where κa becomes very small (small particles with a relatively thick double layer), then Eq. (26.15) reduces to

$$\mu = \frac{2\varepsilon_o\varepsilon\zeta}{3\eta} \tag{26.16}$$

whereas in the other limit, as κa approaches infinity (large particles with very thin double layers), the equation becomes the same as Eq. (26.9). The Smoluchowski equation is derived with similar assumptions as the electro-osmosis mobility equation, and the treatment therefore assumes that the major force retarding the movement of the particles is derived from streaming of the diffuse layer opposite the movement of the particle, also called *electrophoretic retardation*. Hückel's treatment assumed conditions such that the value of κa was small, and the major retarding force was the frictional resistance of the medium rather than electrophoretic retardation. The various conditions can be taken into account by applying a correction factor deduced by Henry which allows Eq. (26.15) to be written as

$$\mu = \frac{2\varepsilon_o\varepsilon\zeta}{3\eta}\left[f\,(\kappa a)\right] \tag{26.17}$$

where $f\,(\kappa a)$ is *Henry's function*. The function $f\,(\kappa a)$ depends on particle shape and, for a spherical particle, varies between 1.0, when $f\,(\kappa a)$ is 0, and 1.5, when $f\,(\kappa a)$ is ∞. Table 26.2 gives some values for Henry's correction function.

This treatment shows that small particles of diverse size and double-layer structure can behave quite differently when conditions are changed just slightly. It is a valuable lesson to see how the size of a particle and its related double layer must

Table 26.2 Selected values for Henry's function

κa	$f(\kappa a)$
0	1.0
1	1.025
10	1.26
100	1.46

be taken into account in understanding the behavior of many molecules of biological importance. Finally, while the dielectric constant and viscosity of the medium must all be taken as equal to the bulk solution values in these derivations, it is likely that this is a gross simplification. For all of these reasons, in studies of electrophoretic mobility and in analytical techniques involving electrophoresis, most researchers use a constant set of conditions and reference their results to those conditions. However, before the role of electrophoretic forces in cellular systems can be quantitatively understood, most of these assumptions would need to be extended.

26.2.5 A Sedimentation Potential Arises When with the Movement of a Particle Relative to a Stationary Solvent

The last of the electrokinetic phenomena is the *sedimentation potential* (*Dorn effect*). Mention of it is included here for the sake of symmetry and completeness. When a particle moves in a stationary solvent, a potential will arise, again due to the separation of the charges, in this case derived from the diffuse layer surrounding the particle. This charge separation occurs because of the substantial difference between the sedimentation coefficient of a macromolecule and the much smaller counterions. Because *s* of the colloidal particle is much larger than that of the ions it will travel much more rapidly leaving the ions behind. The field generated is another force acting counter to the centrifugal force, and the s value is underestimated. Sedimentation potentials are rarely studied, and their measurement and interpretation can be fraught with significant difficulty. Consideration of the effects of sedimentation potentials may be of some importance in ultracentrifugation studies of sedimentation rates especially when experiments are performed in low ionic strength solvents (<0.01 M). In practice the effect is reduced by using ionic strengths of >0.05 M.

26.2.6 Electrokinetic Phenomena Can Have a Role in Biological Systems

Electrokinetic effects are generally considered in the framework of interactions between an electrolyte solution and an insulating surface or in colloidal systems, but since they are derived from the mechanical interactions with the double layer, they may be found acting at any electrified interface including metal–electrolyte systems. Macromolecules and organelles in the cell have many of the characteristics of colloidal systems, and the structure and behavior, for example, of the cytosol of the cell is very much affected by the interphase properties of the electrified interface. The electrokinetic effects are not only restricted to very small systems. For example, the flow of blood past the surfaces of the blood vessel walls is capable of producing a streaming potential. Cells do respond to potentials modeled after these effects but whether this response has a role in the normal development or in pathologic processes of the vessel wall remains to be determined.

26.3 Electron Transfer Is an Essential Form of Biological Charge Transfer

The transport of charge in biological systems when the charge is in the form of ions has been discussed. However charge transfer in many biological systems, such as photosynthesis and in the mitochondrial electron transport chain, is electronic in nature. In our discussion of proton conduction in water we have already seen how the quantum behavior of protons and electrons must be accounted for in these discussions. This lesson will be amplified in the following sections.

26.3.1 Dynamic Electrochemistry Is the Study of Electron Transfer and Their Kinetics

Our attention is focused on electrochemical reactions. Electrochemical reactions involve a net transfer of charge in the overall reaction. The charge transfer usually occurs at the interface between an electronically conducting face and one which is ionically conducting. Thus the charge transfer is interfacial and occurs at an electrode. The physical structure of the interphase is important in modulating the kinetics of the charge transfer. Just as we developed the vocabulary and an understanding of the nature of the interphase by considering the work of physical electrochemists and then relating those ideas to biological systems we use a similar approach studying first previous work on electrochemical kinetics. We will then take the view that the charge transfer processes that occur in biochemical systems may be treated in a similar fashion but in solution: this is the electrodic view of biological electron transfer.

When an electron passes between the electrode and an ionizable species, an electrochemical reaction occurs. Two fundamental differences exist between a chemical and an electrochemical reactions. The first is best described with respect to the potential energy surface of the reactants. While a chemical reaction requires the spatial proximity of the reactants, an electrochemical reaction does not require this proximity. In an electrochemical reaction, the collision of spatially separated reactants may occur by proxy with the intermediary being an electronic conductor. The second difference is in the thermodynamics of these reactions. Electrochemical reactions do not absorb or dissipate heats of reaction. In an ideal reversible electrochemical reaction the heat exchanged is just $T\Delta S$ and the entire free energy of the reaction (at least at low, limiting rates) is available as electrical energy. In a real system in which the rate of an electrochemical reaction is significant, a larger heat than $T\Delta S$ is involved because of the *overpotential*. We will describe the overpotential in the following discussion.

Electrochemical processes are usually studied in an electrochemical cell which is composed of a *working electrode* (*WE*) whose potential is measured with respect to a *reference electrode* (*RE*). Current can be made to flow at the working electrode by connecting a direct current power supply between it and an *auxiliary electrode*. The reference electrode is chosen and maintained so that it maintains a steady potential even when current flows (i.e., it is non-polarizable). If the reference electrode is a normal hydrogen electrode then by convention we define its potential to be 0.00 V. Since the potential at the reference electrode does not vary, all of the changes seen in the system represent changes at the working electrode. We have already considered the thermodynamics of the electrochemical reaction (Chapter 12) and this corresponds to the special condition during which no current flows in the cell. If the cell potential is monitored while no current flow is permitted in the circuit, a steady-state potential value will eventually be reached indicating the arrival at equilibrium. The potential of the working electrode at equilibrium is given by Nernst equation:

$$E_{\mathrm{WE}} = E^{o} + \frac{RT}{nF} \ln \frac{c_{\mathrm{ox}}^{\sigma}}{c_{\mathrm{red}}^{\sigma}} \tag{26.18}$$

The potential is generally written in terms of the standard cell potential and the surface concentrations (c^{σ}) of the redox couple. (This expression assumes that the activity coefficient is unity so that the concentrations may be used in place of activities.) The standard potential is the equilibrium potential that is found when the concentrations of the redox pair are equal. Since no current flows, no electrochemical reaction can occur and the bulk concentration (c^{∞}) of the redox couple must be equal to the surface concentrations. Thus the measured potential at the working electrode is the same as the cell potential:

$$E_{\mathrm{CELL}} = E_{\mathrm{WE}} - E_{\mathrm{RE}} = K_{\mathrm{WE}} - 0 \tag{26.19}$$

$$E_{\mathrm{CELL}} = E^{o} + \frac{RT}{nF} \ln \frac{c_{\mathrm{ox}}^{\infty}}{c_{\mathrm{red}}^{\infty}} \tag{26.20}$$

We know that equilibrium really represents a dynamic steady state. There is current flow at the surface of the electrode in which the rate of oxidative charge transfer equals the rate of reductive charge transfer:

$$-\overrightarrow{I} = \overleftarrow{I} = I_0 \tag{26.21}$$

Here the *partial current densities* of the forward (cathodic) reaction and the back (anodic) reaction are $-\overrightarrow{I}$ and $\overleftarrow{I}$, respectively. The cathodic current is given a negative sign by convention and when the symbol appears the value is usually understood to be negative. The term I_0 is called the *exchange current density* and is an important kinetic parameter in the electron transfer process. The power supply can be adjusted so that the potential of the WE electrode is made more negative than the equilibrium potential. This will force the cell to seek a new equilibrium by re-establishing new surface concentrations of the redox couple in accordance with the Nernst equation. To change the concentrations of the redox couples at the electrode, a certain amount of current must flow in the cell causing the electrochemical conversion of the oxidized species into reduced species. The converse process can be described if the working electrode is made positive. The linkage between the current and potential is called the *I–E* behavior. When current flows in an electrochemical cell, the overall circuit must obey Kirchoff's laws. The applied potential will be dropped across three circuit elements:

$$E = E_{\mathrm{eq}} + IR + \eta \tag{26.22}$$

E_{eq} is the equilibrium potential as we have already discussed. The *IR* term represents a potential drop secondary to the resistance (*R*) found between the two electrodes in the system and η is the *overpotential*. The resistance of the electrolyte can usually be made insignificant by physically approximating the components of the cell closely. The overpotential can then be defined as

$$\eta = E - E_{\mathrm{eq}} \tag{26.23}$$

The overpotential is the electrical potential beyond the equilibrium potential necessary to make current flow in a real cell. In a sense it represents the free energy difference between the ideal reversible system and the real irreversible system. Thus the overpotential is a measure of a given system's deviation from ideality.

So far we have described the *I–E* behavior of a cell in terms of thermodynamics. But the flow of current at a given potential depends on the kinetics of the electron transfer and the rate of mass transfer of the reactants to the electrodic surface. The overall current density is

$$\overrightarrow{I} + \overleftarrow{I} = I \tag{26.24}$$

The partial current densities depend on the concentration of the electroactive species at the electrode and a rate constant:

$$\overrightarrow{I} = nF\overrightarrow{k}c_{\mathrm{ox}}^{\sigma} \tag{26.25}$$

$$\overleftarrow{I} = nF\overleftarrow{k}c_{\mathrm{red}}^{\sigma} \tag{26.26}$$

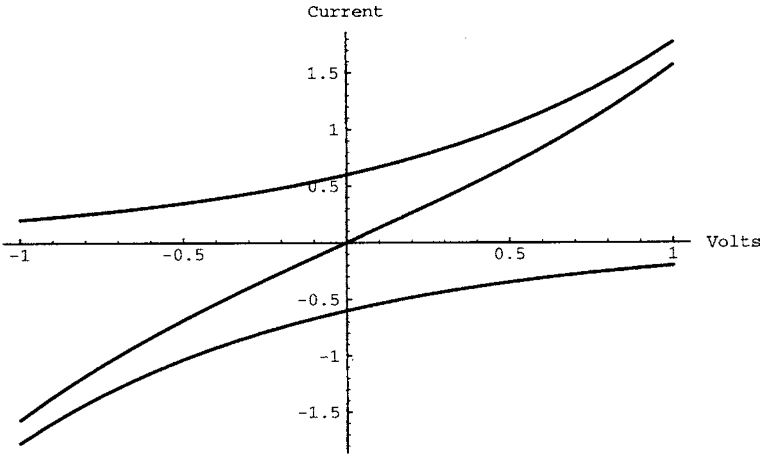


Fig. 26.1 *I* versus *E* curves for an electrodic system. The kinetic constants I_0 and α can be determined from a plot of this type

Experimentally, it is found that the rate constants are dependent on the electrode potential (Fig. 26.1) and these can be written as

$$\overrightarrow{k} = \overrightarrow{k}_o \exp\left(\frac{-\alpha_\mathrm{C} nF}{RT}E\right) \tag{26.27}$$

$$\overleftarrow{k} = \overleftarrow{k}_o \exp\left(\frac{\alpha_\mathrm{A} nF}{RT}E\right) \tag{26.28}$$

The terms α_A and α_C are constants (usually valued at 0.5) called the *transfer coefficients* for the anodic and cathodic reactions and for simple electron transfer reactions $\alpha_\mathrm{C} + \alpha_\mathrm{A} = 1$. The k_o terms are the rate constants at the standard potential E^o.

With a little algebra to combine the relationships of the overpotential (Eq. 26.9) and the current density relationships, we can derive the *Butler–Volmer* equation.

$$I = I_\mathrm{o}\left[\exp\left(\frac{\alpha_\mathrm{A} nF}{RT}\eta\right) - \exp\left(\frac{-\alpha_\mathrm{C} nF}{RT}\eta\right)\right] \tag{26.29}$$

This equation is the fundamental equation of electrode kinetics and describes the linkages between the current density, and the exchange current densities, the overpotential, and the transfer coefficients. The Butler–Volmer equation is often written in simplified form; in the cases when the overpotential is either very high or very low. These conditions limit the Butler–Volmer equation and are called the *Tafel* equations. Their derivation is left as an exercise to the student. Equation (26.29) is the most general form of the Butler–Volmer equation and the transfer coefficients are valid for multi-step reactions including those with electron transfer steps other than the rate determining step of the overall reaction.

The transfer coefficient determines how the input of electrical energy affects the reaction rate. If we consider a much simpler single-step, single-electron transfer, the transfer coefficient is sometimes called a *symmetry factor* β. The symmetry factor can be physically interpreted in terms of a potential energy curve that connects the reaction path of the initial, activated, and final states. Thus for a de-electronation (oxidation) or electronation (reduction) reaction, the electron must move along a potential energy surface whose dimensions represent the interplay between the chemical reactants and the electric field. Obviously β will reflect to some degree the structure of the double layer and the properties of the reacting and final species involved in the reaction.

26.3.2 Electron Transfer Is a Quantum Mechanical Phenomenon

We will pursue this idea of the reaction path of an electron moving along a potential energy surface from reactants (half a red*ox* pair plus the electron) to products (the other half of the *red*ox pair). Though this is written as a reduction reaction but clearly an oxidation reaction could just as well be considered. Why do changes in the electrode potential affect $\overrightarrow{k}$ and $\overleftarrow{k}$? We will limit our discussion to a simple one electron transfer. The reaction can be written in the generalized form:

$$O + e^- \rightleftharpoons R \tag{26.30}$$

The product and reactant are each represented by a well in the potential energy surface separated by a saddle point whose height is determined by the transition state energy barrier (Fig. 26.2). The system must move through the energy barrier for an electron transfer to take place in either direction. At equilibrium the transition rates across the barrier are equal in the forward and reverse directions and $\overrightarrow{k} = \overleftarrow{k} = k^o$. Using absolute rate theory to relate the activation energy and the rate constants we can write

$$k = KZe^{-\Delta G^*/RT} \tag{26.31}$$

K is the transmission coefficient and Z is given by $\frac{k_b T\sigma}{h}$, σ is a reaction length that is approximately a molecular diameter. When the concentrations of the oxidized and reduced species are equal, $\Delta\overrightarrow{G}^{o*} = \Delta\overleftarrow{G}^{o*} = \Delta G^{o*}$. Thus the activation energy barrier at equilibrium must be symmetrical. The species involved in electrochemical reactions are sensitive to the Galvani potential difference across the interface and this sensitivity is reflected in the associated $\Delta\overrightarrow{G}^*$ and $\Delta\overleftarrow{G}^*$. When the electrode potential is moved from an equilibrium value, the free energy curves for O and R will move apart in the y-dimension by

$$\Delta G = nF(E - E^o) \tag{26.32}$$

The reaction path is now asymmetric (Fig. 26.2b) and a net reaction will occur.

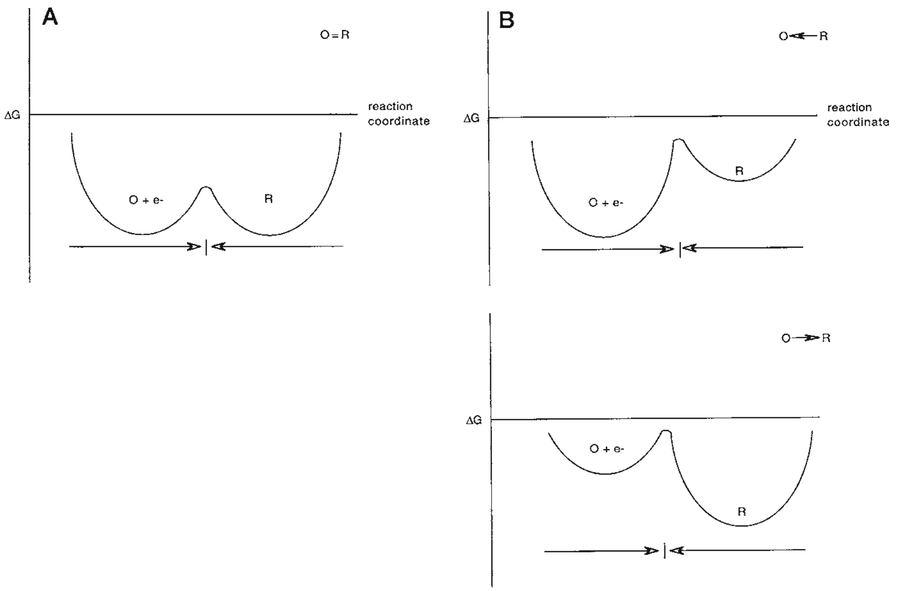


Fig. 26.2 Potential energy curve for the course of a simplified electrochemical reaction. (**a**) At equilibrium the reaction path is symmetrical around the transition point; (**b**) a change in electrode potential causes the potential energy curve to be asymmetric and a net reaction will occur

In order to proceed quantitatively to the calculation of the rates of the forward and reverse reactions, knowledge of the shape of the energy curves is required. For simplicity our analysis will use parabolic curves thus assuming harmonic motion. In this case it can be shown that $\Delta \overrightarrow{G}^* = \Delta \overleftarrow{G}^*$ are related to ΔG^{o*} as

$$\Delta \overrightarrow{G}^* = \Delta G^{o^*} + \frac{n^2F^2\,(E-E^o)^2}{16\Delta G^{o*}} + \frac{1}{2}nF\left(E-E^o\right) \tag{26.33}$$

$$\Delta \overleftarrow{G}^* = \Delta G^{o^*} + \frac{n^2F^2\,(E-E^o)^2}{16\Delta G^{o*}} - \frac{1}{2}nF\left(E-E^o\right) \tag{26.34}$$

The free energies of activation can be related to the rate constants for the forward and reverse reactions by insertion into Eq. 26.31) and the potential dependence of $\overrightarrow{k}$ and $\overleftarrow{k}$ can be written as

$$\overrightarrow{k} = k^o \exp\left(\frac{\frac{n^2F^2(E-E^o)^2}{16\Delta G^{o*}} - \frac{1}{2}nF\,(E-E^o)}{RT}\right) \tag{26.35}$$

$$\overleftarrow{k} = k^o \exp\left(\frac{\frac{n^2F^2(E-E^o)^2}{16\Delta G^{o*}} + \frac{1}{2}nF\,(E-E^o)}{RT}\right) \tag{26.36}$$

Though these equations appear complex they can be made familiar by recognizing that they are the same as Eqs. (26.27) and (26.28) because

$$\alpha_c = \frac{1}{2} + \frac{nF\,(E - E^o)}{16\Delta G^{o*}} \tag{26.37}$$

Thus the relative position of the transition point with respect to the spatial dimensions of the potential energy curves varies with the electrode potential (Fig. 26.2b). It should apparent that varying the concentrations of the reduced and oxidized species will also alter the relative positions of the curves with respect to the transition point. The manipulation of the potential energy surface by the electrode potential and the redox ratio and its reflection in the kinetic rates of the forward and back reactions has now been given a physical interpretation.

We can write the symmetry factor in the following terms:

$$\beta = \frac{\text{distance along reaction coordinate between initial and transition state}}{\text{distance along reaction path between initial and finalstate}}$$

Alternatively if we consider the movement from reactant to product as requiring the stretching of the bonds along the reaction path, then the electrochemical potential represents a source of electrical work that can be introduced into the reaction and alters the activation energy.

$$\beta = \frac{\Delta\ \text{activation energy}}{\Delta\ \text{electrical energy}} = \frac{\Delta U^*}{\Delta F\eta} \tag{26.38}$$

By using geometric arguments, β can be related to the slopes of the potential energy curves leading to the transition states. Such an analysis is available in the references cited at the end of the chapter. The description of the reaction in terms of a potential energy curve emphasizes the dependence of the charge transfer reaction on the nuclear position of the reactants and the products. In most electrochemical reactions the actual charge transfer is electronic. We know the motion of the electron must be treated quantum mechanically. Proton (H^+) and hydride (H^-) transfer also play a role in charge transfer reactions especially in biological systems. Comparing the de Broglie wavelengths of the major charge transfer species in biological systems we can see that it is the electronic and protonic charge transfers that must be viewed with a quantum perspective. We recall from our earlier discussions that because of the long tail of their wavefunctions, quantum mechanics allows tunneling through a classical energy barrier such as the activation energy barrier between the reactant and product species. From a kinetic standpoint the probability of tunneling through a barrier is not in itself sufficient to account for the observed rate because the rate is also dependent on the presence of accessible landing sites on the other side of the barrier. We have now come to the point where we must treat the charge transfer process as dependent on both nuclear position and electron tunneling.

Up to this point we have been regarding heterogeneous charge transfer at an electrode–solution interface. It is clear that the energy level of the donor electron and the energy level of the empty receptor site must be considered. Metals are relatively straightforward because the electron energy levels are treated as delocalized energy bands created by the diffraction of the electron waves by the crystal lattice. The population of these bands is then determined by a Fermi–Dirac distribution function which leads to a sharply formed distribution with a maximum energy EF called the Fermi energy. Semiconductors distribute the energy levels into two relatively distinct regions, a valance band and a conduction band. Valence band electrons are non-conductive and electrons promoted into the conduction band are free to move. The electronic energy levels of ions and molecules in solution are different because they are localized. The energy of the electronic states involved, both the jump site and receptor site, vary rapidly over time with response to the local fluctuating polarizing forces in the solution. We use a normal distribution function (Gaussian) with a peak at the most probable energy to describe the overall state of the solution-based system. There will be two separate normal populations described in the solution for the redox pair because the charge distribution of the oxidized and reduced species are distinctly different. The location of the Fermi levels for each of these general systems are shown in Fig. 26.3. It is relatively easy to imagine that the solution-based distribution system in contact with a metal will come to equilibrium by the transfer of electrons from occupied orbitals on one side of the interface tunneling across the interface to unoccupied orbitals of equivalent energy. At equilibrium there will be a free exchange back and forth across the interface between these degenerate levels.

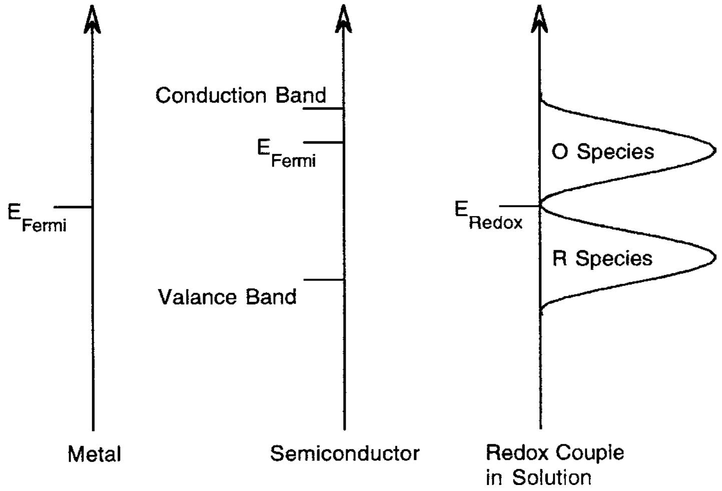


Fig. 26.3 Energy level diagram descriptions for metal, semiconductor, solution-based electrochemical systems

26.3.3 Electron Charge Transfer Can Occur in Proteins

Electron transfer reactions such as those found in photosynthetic centers or in the mitochondrial electron transport chain are generally treated as a charge transfers between redox centers in a solvent medium. Thus the protein is either treated as a structureless continuum or as a solvent of specific structure that modulates the charge transfer reaction. This question defines the current state of the art and is an area of active research interest. Before proceeding, we must first ask: What is the molecular picture of the charge transfer process in solution? The picture in broad strokes is that the rate of electron transfer depends on two factors, one nuclear and one electronic:

1) Nuclear factor – Electron transfer will occur only under the condition that the donor and acceptor orbitals are the same energy, i.e., degenerate. The reactant and product systems each rest on a potential energy surface that defines their nuclear configuration. This configuration determines the energy state of the electronic orbitals. The energy levels of importance are those electronic levels between which charge transfer will occur.
2) Electronic factor – The electron transfer occurs via tunneling through the potential energy barrier of the transition state. This part of the process depends on the interaction of the electronic wavefunctions between the donor and acceptor redox species. The larger the overlap of the wavefunctions, the more likely is the probability of observing the tunneling event.

In other words, electron transfer is dependent on a condition in which certain energy states are accessible and the rate of transfer is related to the probability of finding an electron in one of those accessible states. We have alluded to this type of system earlier when we discussed thermodynamics and made reference to the quantum mechanical construct called Fermi’s Golden Rule (Appendix Q). The starting point for a mathematical expression reflecting the kinetics of electron transfer is made by using this rule with the additional condition that electron transfer is a non-adiabatic process (Appendix R).

$$k_{\mathrm{ET}} = \frac{4\pi^2}{h} V^2 \mathrm{FC} \tag{26.39}$$

V represents the electronic component and is the electronic coupling matrix element. FC is the nuclear component in which the Franck–Condon principle plays a central role; it is called the Franck–Condon factor. FC actually represents the overlap of the vibrational wavefunctions of the reactant and product states.

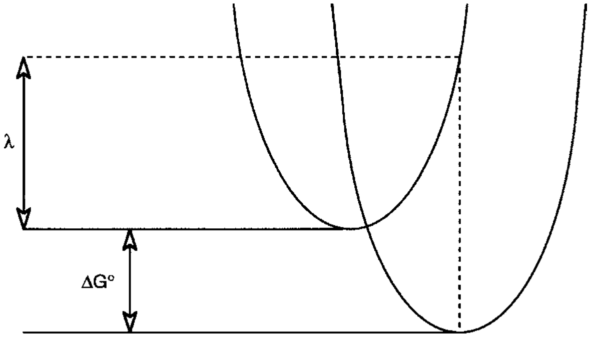
We can provide a somewhat more detailed description by starting with the Born–Oppenheimer approximation. The argument is made that the electronic motion is so much faster than the nuclear motion, and therefore, that the nuclear motion and electronic motion may be separated. Two Hamiltonians can be written as

$$H\psi_{\mathrm{e}} = U\sum e \text{ (electrons, fast system)} \tag{26.40}$$

$$H\psi_{\mathrm{n}} = U\sum n \text{ (nuclei, slow system)} \tag{26.41}$$

First, we will consider the nuclear component. Figure 26.4 shows how we conceive of the overall potential energy curve as shaped by two harmonic functions, one each for the reactant and the product. The energy separating the two minimums is $\Delta G_{\text{reaction}}$ and the intersection of the two curves above the point of minimum energy defines the geometry of the transition barrier. When the energies of these two subsystems become degenerate, electron transfer can occur between an occupied and unoccupied orbital via tunneling. These become degenerate by the distortion of the nuclear product state into the same energy state as the reactant state. The vibrational energy levels for each subsystem are described in terms of their quantum mechanical wavefunctions. Because the electron transfer step is very fast compared to the nuclear movements the Franck–Condon principle can be applied to find the vibrational energy configurations that lead to overlap of the vibrational wavefunctions of the two nuclear potential energy curves. The point of degeneracy and hence of electron transfer is described in terms of a Franck–Condon factor.

Fig. 26.4 Potential energy curve for the nuclear component of the electron transfer step. The reactant and product states are represented by harmonic functions with electronic energy levels shown for each state. The geometry for the discussion in the text is shown here



The calculations of the Franck–Condon factor are based on the classical calculation by Marcus and the quantum mechanical refinements proposed by Dogonadze and Levich and by Jortner. In this model, the transition state leading to charge transfer is reached because of the thermal fluctuations of the local solvent and/or the ligands contacting the redox-active species. The theory treats charge transfer in reactions in which no bonds are broken or formed and does not take into account the ionic atmosphere's contribution to the energy of activation. With these substantial simplifying assumptions explicitly stated we proceed with a description of the charge transfer process. In this context the relative positions of the electrochemical system consisting of the reactant and product in nuclear state space along with their surroundings (solvent, ligands, ionic atmosphere, etc.) are described in terms of a reorganization energy, λ. Marcus derived an expression for λ using a modeling method of "hard spheres in a dielectric continuum" similar to the one that we have already explored in our analysis of ion solvation. The reorganization energy represents the energy needed to distort the product state into the same shape as the reactant state without electron transfer occurring (Fig. 26.5). When the fluctuations of the two reacting molecules are such that the two nuclear states are identical, their electronic energy levels will be degenerate and electron transfer can occur. The form of the FC factor as derived by Marcus is written:

$$\mathrm{FC} = (4\pi\lambda kt)\,\exp\left[{}^{-(\Delta G^{o}+\lambda)^{2}}/{}_{4\lambda kt}\right] \tag{26.42}$$

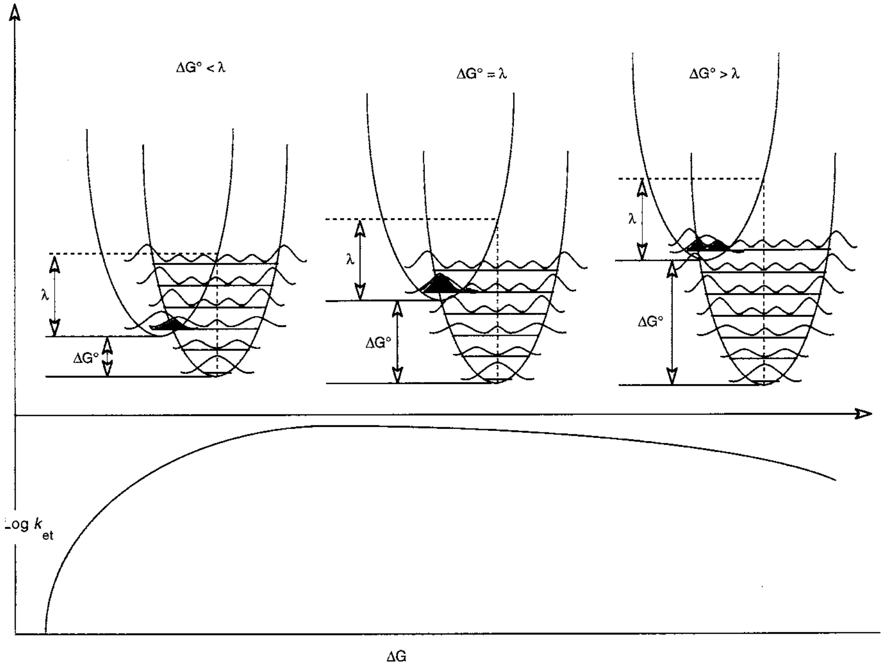


Fig. 26.5 (*Upper*) Consideration of the Franck–Condon principle allows demonstration of the geometry of interaction between the reactant and product curves leading to the conditions of maximal FC factor. (*lower*) Plot of k_{et} inversion as a result of the varying reorganizational energy with respect to $\Delta G°$

The important property of λ, that the FC factor (and hence k_{et}) has a maximum when $\Delta G^{o} = \lambda$ as shown in Fig. 26.5, is noted in this expression. This inverted maximum can be understood by considering the intersection of the degenerate energy levels consistent with the Franck–Condon principle, which makes the most likely transition the one with the greatest vibrational overlap. In Fig. 26.5 we can see why the level corresponding to the lowest energy in the reactant state has maximal overlap with a degenerate vibrational energy level in the product state when $\Delta G^{\text{o}} = \lambda$. In this figure, λ is kept constant and ΔG^{o} is altered. The inverted effect has been convincingly demonstrated by measurements of electron transfer rates in both synthetic and biological systems.

In polar solvents the major contribution to λ is the reorientation of the solvent molecules that results from the change in charge distribution of the reactant. The calculation of the solvent reorganization energy can be made with a dielectric continuum model. The simplest models treat the reactant as a conducting sphere. More complicated models treat the reactant inside a cavity (of variable geometry) of low dielectric strength. If a reactant is placed inside a low dielectric medium such as the bilipid membrane, the reorganization energy can be substantially lowered. Further, λ will generally rise with increasing polarity of the solvent. A second component of the reorganization energy derives from alterations in the geometry of reactant as it becomes product. These changes in bond length and angle are due to the redistribution of charge that occurs with the charge transfer. These changes are inherently quantum mechanical and in general classical approximations of these "inner-sphere" reorganization are inadequate so quantum mechanical treatments are appropriate.

The electronic component of the electron transfer rate is characterized by a non-adiabatic interaction in which there is a weak interaction between the reactants and products at the transition state configuration. The electronic factor, V_R^2, depends on the coupling strength, (V_o^2) the edge-to-edge distance between the reactant and product (R) and a coefficient of decay of the coupling constant (β) with respect to R:

$$V_R^2 = V_o^2 e^{-\beta R} \tag{26.43}$$

The electron transfer rate will fall off exponentially with respect to distance. In this somewhat simplified model, the electron transfer step is treated as an electron tunneling event through a square well potential. The height of the barrier with respect to the energy states of reactant–product interaction determines β, which defines the distance-decay factor. In a vacuum β has a value of 3.5–5 Å^{-1}, thus making the rate of electron transfer less than 10 s^{-1} at R of greater than 8 Å. For all practical purposes this rate is prohibitively slow. If a medium is placed between the reactant–donor pair, the height of the tunneling barrier is reduced leading to a lower value for β. Hopefield estimated that β would be 1.4 Å^{-1} for an electron transfer through a protein matrix corresponding to a decrement in k_{et} of 10-fold per 1.7 Å increase in R. This means that the interaction and the electron transfer rate between the reactant and product is substantially larger when there is an intervening protein compared with electron transfer with free space separating the donor–acceptor (DA) pair.

The lower value of β in the presence of a protein occurs because the interaction of the electronic charge distribution in the biological matrix and the DA wavefunctions effectively increases the amplitude of the tunneling wavefunction and increases the likelihood of a charge transfer event occurring in a given time thus increasing the rate. What is the best way to handle this interaction between interacting wavefunctions and the increased coupling of these interactions by intervening matrix? On one hand, the biological matrix can be treated as a continuum in which essentially every path of the ET step is equivalent. In this view proteins can be treated as an organic glass and detailed knowledge of the intervening protein structure is not necessary. This is the position taken by the group of PL Dutton who argue that a single value of β is a fundamental characteristic of proteins in general. This viewpoint suggests that the control of electron transfer kinetics in biological systems can be fully provided by alterations in ΔG^{o} and λ (over a range of 10^5 fold) and most importantly by distance (providing a range of 10^{12} fold in rate, see Fig. 26.6).

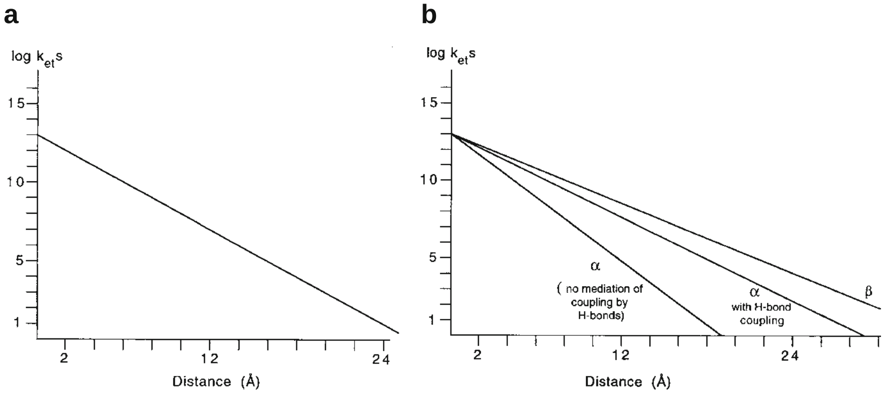


Fig. 26.6 (*Left*) Plot of electron transfer rate as a function of distance in the photosyntheic reaction center (data from Moser et al. (1992) *Nature*, 355:796). (*Right*) Plot of k_{et} versus distance for ruthenium-modified proteins with different secondary structures coupling the donor–acceptor pairs (data from Gray and Winkler (1996), *Annu. Rev. Biochem.*, 65:537)

An alternative view is supported by the work of HB Gray's group who have used ruthenium-modified proteins to investigate the coupling between the donor and acceptor pair. The coupling interaction is not directly between the DA pair but rather through an intervening bridge of covalent and hydrogen bonds that couples the exchange. There is a specific bridge (or series of bridges) composed from covalent and hydrogen bonds and jumps in free space. A structure-dependent tunneling pathway can be proposed linking the donor to the acceptor. Computational techniques that use a structure-dependent searching algorithm can be used to find the optimal pathway. One of the results of this approach is that proteins dominated by β-pleated sheets (β of 0.9–1.15 Å^{-1}) are better a mediating long-range coupling than those built predominantly from α-helices (β of 1.25–1.6 Å^{-1}) (Fig. 26.6). The tunneling-pathway model of electron transfer supports the notion that secondary structure in a protein specific fashion mediates the coupling efficiency of the transfer rate.

The control of electron transfer rate can be effected by both nuclear and electronic factors in biological systems. Distance clearly plays an important role and the overall tertiary and quaternary structure of a protein will effect such control. Placement of redox active sites in a charge relay system will allow substantial control of the electron transfer process. Intervening secondary structure likely plays an important role in some protein systems. In terms of the nuclear factors, the relatively limited values that ΔG^o can take in biological systems are less likely to play a major role in control when compared with the ΔG^o–λ interaction. The inverted region behavior can be used in biological systems to control the direction of electron transfer flow. For example, the rapid one-way flow of electrons in the photosynthetic center of *Rhodopseudomonas viridis* which generates the initial charge separation proceeds at almost 100% quantum efficiency even through the back reaction of BPh^- to BChl_2 is actually favored over the reaction BPh^- to Q_A from the thermodynamic standpoint. Figure 26.7 shows the arrangement of the chromophores in this photosynthetic system. The measured ET rates are indicated in this figure. The initial charge separation occurs with a relatively large photoelectronic energy of 1.4 eV acting in a system with small λ and ΔG^o. The large $-\Delta G^o$ of the back reaction pushes the back reaction into the inverted region which effectively retards the rate of the back reaction from BPh^- to BChl_2 thus assuring the productive

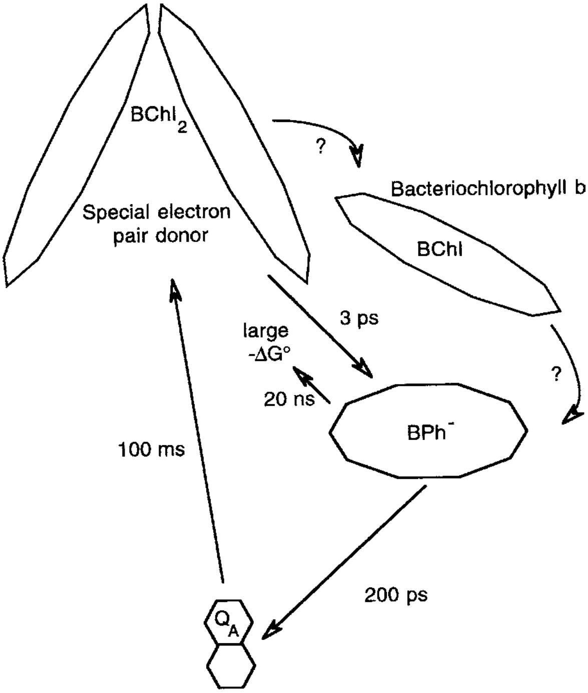


Fig. 26.7 Chromophore arrangement of the *Rps. viridis* photosynthetic center. Note the rates of ET at room temperature noted on the diagram (data from Boxer (1990) *Annu. Rev. Biophys. Biophys. Chem.*, 19:267)

reaction of BPh^{-} to QA . This nuclear effect dominates over the electronic effects since a similar distance is found between these two redox centers.

Further Reading

General Texts

Hunter R.J. (1981) *Zeta Potential in Colloid Science. Principles and Applications.* Academic Press, London.

Dynamic Electrochemistry

General

Albery W.J. (1980) The application of the Marcus relation to reactions in solution, *Annu. Rev. Phys. Chem.*, **31**:227–263.

Bockris J.O'M., Reddy A.K.N., and Gamboa-Aldeco M. (1998), *Modern Electrochemistry*, 2nd edition, Volume 2, Electrodics. Kluwer Academic/Plenum, New York.

Faulkner L.R. (1983) Understanding electrochemistry: some distinctive concepts, *J. Chem. Educ.,* **60**:262–264.

Khan S.U., Kainthla R.C. and Bockris J.O'M (1987) The redox potential and the Fermi level in solution, *J. Phys. Chem.,* **91**:5974–5977.

Voltammetry

Evans D.H., O'Connell K.M., Petersen R.A., and Kelly M.J. (1983) Cyclic voltammetry, *J. Chem. Educ.*, **60**:290–298.

Kissinger P.T. and Heineman W.R. (1983) Cyclic voltammetry, *J. Chem. Educ.*, **60**:702–706. (The fundamentals with some experimental data describing some biological compounds of interest.)

Biological Electron Transfer

Barbara P.F., Meyer T.J., and Ratner M.A. (1996) Contemporary issues in electron transfer research. *J. Phys. Chem.*, **100**:13148–13168.

Beratan D.N., Betts J.N., and Onuchic J.N. (1991) Protein electron tunneling transfer rates set by the bridging secondary and tertiary structure, *Science*, **252**:1285–1288.

Beratan D.N., Onuchic J.N., Winkler J.R., and Gray H.B. (1992) Electron tunneling paths in proteins, *Science*, **258**:1740–1741.

Boxer S.G. (1990) Mechanisms of long-distance electron transfer in proteins: lessons from photosynthetic reaction centers, *Annu. Rev. Biophys. Biophys. Chem.*, **19**:267–299.

Gray H.B. and Winkler J.R. (1996) Electron transfer in proteins, *Annu. Rev. Biochem.*, **65**:537–561.

Gray H.B. and Winkler J.R. (2003) Electron tunneling through proteins. *Quart. Rev. Biophys.*, **36**:341–372.

Gray H.B. and Winkler J.R. (2005) Long-range electron transfer. *PNAS*, **102**:3534–3539.

Khan S.U. (1988) Models of electron transfer reactions at a biological-membrane covered electrode-solution interface, *J. Phys. Chem.*, **92**:2541–2546.

Kuki A. and Wolynes P.G. (1987) Electron tunneling paths in proteins, *Science*, **236**:1647–1652.

Marcus R.A. (1993) Electron transfer reactions in chemistry: theory and experiment, *Rev. Modern Phys.*, **65**:599–610.

McMahon B.H., Joachim D., Müller J.D., Wraight C.A., and Nienhaus G.U. (1998) Electron transfer and protein dynamics in the photosynthetic reaction center, *Biophys. J.*, **74**:2567–2587.

Moser C.C., Keske J.M., Warncke K., Farid R.S., and Dutton P.L. (1992) Nature of biological electron transfer, *Nature*, **355**:796–802.

Muendon O. and Hake R. (1992) Interprotein electron transfer. *Chem. Rev*., **92**:481–490.

Onuchic J.N., Beratan D.N., Winkler J.R., and Gray H.B. (1992) Pathway analysis of protein electron-transfer reactions, *Annu. Rev. Biophys. Biomol. Struct.*, **21:**349–377.

Pascher T., Chesick J.P., Winkler J.R., and Gray H.B. (1996) Protein folding triggered by electron transfer, *Science* **271**:1558–1560.

Problem Sets

1. The blood vessel can be modeled as a tube with an electrolyte traveling through it under a hydrostatic pressure. The ζ potential of the vessel is –200 mV; the pressure is 9.4×10^3 N/m^2; blood viscocity is 3.5×10^{-3} kg/m/s; the conductivity is 0.67 S/m; use a dielectric constant of 10.

   (a) What is the streaming potential under these conditions?
   (b) What happens to the streaming potential if the patient is hypertensive and the mean arterial blood pressure doubles?
   (c) What assumptions are made in this analysis?
   (d) Are these assumptions valid?

2. The zeta potential depends on the surface charge of the vessel. Endothelial cells have a net negative surface charge. However in cases of vascular injury the endothelial cells are stripped off to reveal the sub-endothelial lining which contains a high content of collagen and elastin. The p*I's* of these proteins can be found by consulting the isoelectric focusing gel shown. Given the blood and tissue pH of 7.4, what happens to the streaming potential with endothelial injury?
3. Explain how electroneutrality is maintained in an electrical circuit that has an electrolytic conductor as a circuit element. Be sure to address the three general aspects of the circuit's behavior, that is, electronic, ionic, and electrodic.
4. Derive the Tafel equations from the Butler–Volmer equation.

Part V
Methods for the Measuring Structure and Function

Chapter 27
Separation and Characterization of Biomolecules Based on Macroscopic Properties

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_27, © Springer Science+Business Media, LLC 2010

27.1 Introduction: Mechanical Motion Interacts with Mass, Shape, Charge, and Phase to Allow Analysis of Macromolecular Structure

In Chapter 2 the importance of identifying the components in a system was introduced. A component may be distinguished because of certain characteristic properties that allow them to be separated from each other and then uniquely identified. Among the most useful macroscopic properties of a specific element within a system are its characteristic motions when subjected to either a constant or an accelerating force field. This often allows both separation and concurrent characterization. There are a variety of motions that are commonly encountered in biophysical separations. We will briefly describe several of these; first motion affected by normal, frictional, buoyant, and drag forces and then methods of inducing motion by means of applying acceleration forces.

27.2 Buoyant Forces Are the Result of Displacement of the Medium by an Object

An object suspended in fluid is weighed with a spring scale and is found to have a smaller weight than when compared to its suspension in air. An upward force called the *buoyant force* is acting on the object. The buoyant force is more dramatically shown when an inflated balloon is submerged in the water: it will be accelerated upward to the surface if allowed to move freely after being submerged. The buoyant force depends on the density of the fluid and the volume of the body but not on the composition or shape of the body. The force is equal in magnitude to the weight of the fluid displaced by the body. This principle is known as *Archimedes principle*. This principle can be derived from Newton’s laws with the result written:

$$F_{\text{buoyant}} = w_{\text{o}} = w_{\text{displaced fluid}} \tag{27.1}$$

where w_{o} is the weight of the object. Whether an object sinks or floats in the fluid depends on the density of the object relative to that of the fluid. ρ_{f} is the density of the fluid. A volume V of the fluid has a mass of $\rho_{\text{f}}V$ and a weight equal to $\rho_{\text{f}}gV$. The weight of the object depends on its density and is $\rho_{\text{o}}gV$. When $\rho_{\text{o}} < \rho_{\text{f}}$ the buoyant force will accelerate the object to the top of the fluid. At equilibrium, it will float with a fraction of its volume submerged so $w_{\text{o}} = w_{\text{f}}$. As we will discover in the following sections, the practice of centrifugation and density gradient centrifugation is direct results of Archimedes principle.

27.2.1 Motion Through a Medium Results in a Retarding Force Proportional to Speed

In addition to buoyant forces, an object moving through a fluid (including an atmosphere) will experience retarding forces that are proportional (to a reasonable approximation) to the velocity of the object:

$$F_\mathrm{r} = -bv \tag{27.2}$$

The proportionality constant, b, depends on the size and shape of the object and on the fluid. Consider the motion of a particle dropped from rest and under the influence of its own weight and a retarding force but neglecting buoyancy. (Buoyancy can be easily added by writing the weight of the object as w–B, where B is the buoyant force.) The downward force is taken positive and the retarding force is upward. From $\sum \mathbf{F} = m\boldsymbol{a}$ we write

$$mg - bv = ma = m\frac{dv}{dt} \tag{27.3}$$

thus

$$\frac{dv}{dt} = g - \frac{b}{m}v \tag{27.4}$$

This has the following solution:

$$v = \frac{mg}{b}\left(1 - e^{-bt/m}\right) \tag{27.5}$$

This differential equation describes motion in which the body initially accelerates quickly, but as the velocity rapidly increases so does the retarding force. At a particular speed, the *terminal velocity*, the retarding force is equal to the weight mg and the acceleration falls to zero. At terminal velocity the body moves with a constant velocity.

Terminal velocity is found by setting ${}^{dv}\!/_{dt} = 0$ which gives the result:

$$mg - bv_t = 0$$

$$v_t = \frac{mg}{b} \tag{27.6}$$

Equation (27.5) can be written in terms of the terminal velocity:

$$v = v_t\left(1 - e^{t/\tau}\right) \tag{27.7}$$

The time constant τ is equal to ${}^{m}\!/_{b}$ and is the time in which the particle reaches 63% of its terminal velocity. As b increases, the terminal velocity decreases and the time necessary to reach a significant percentage of the terminal velocity also falls. b depends on the shape of the object and these considerations can give information about the physical properties of molecules.

27.2.2 Frictional Coefficients Can Be Used in the Analysis of Macromolecular Structure

The frictional force experienced by an object such as a macromolecule depends on the velocity of the particle, v_{i}, and a constant, f_i, the *frictional coefficient*. The frictional coefficient depends on the size and shape of the object and can be determined by applying a known force, $\mathbf{F}_i$, to a solution of the molecule under study. A constant velocity is reached when the total force on the molecule is 0:

$$f_i v_i - \mathbf{F}_i = 0 \tag{27.8}$$

A sphere is found to have the minimum frictional coefficient called f_{o}. The exact solution for the frictional coefficient for a sphere of radius r is given by Stokes law:

$$f_{\mathrm{o}} = 6\pi\eta r \tag{27.9}$$

where η is the viscosity of the medium. Other shapes can be related to the spherical coefficient, f_{o}, in the form of a ratio $\frac{f_i}{f_{\mathrm{o}}}$ that is always larger than 1. The ratios for the ellipsoid and rod-like shapes are given in Table 27.1. As Fig. 27.1 shows, the shape and size of a particle or macromolecule give a distinct contour when the axial ratio is plotted against $\frac{f_i}{f_{\mathrm{o}}}$. The frictional coefficient depends on the linear dimensions of particles and provides information about the particle's size and shape in

Table 27.1 Frictional coefficients

	$\frac{f_i}{f_{\mathrm{o}}}$	R_{e}
Ellipsoid (prolate)	$\frac{P^{1/2}(P^2-1)^{1/2}}{\ln\left[P+(P^2-1)^{1/2}\right]}$	$\left(ab^2\right)^{1/3}$
Ellipsoid (oblate)	$\frac{(P^2-1)^{1/2}}{P^{1/3}\tan^{-1}(P^2-1)^{1/2}}$	$\left(a^2b\right)^{1/3}$
Rod	$\frac{\left(\frac{2}{3}\right)^{1/3}P^{1/3}}{(\ln 2-0.3)}$	$\left(\frac{3a^2b}{2}\right)^{1/3}$

The frictional ratio is written in terms of the geometries of the non-spherical shapes. The semi-major axis is a and the axial ratio defining the spheroid is $\frac{a}{b}$. The axial ratio for a sphere is 1. We designate the axial ratio as P. a is the half-length measure for a rod. Here R_{e} gives the radius of a sphere whose volume is equal to the rod or ellipsoid

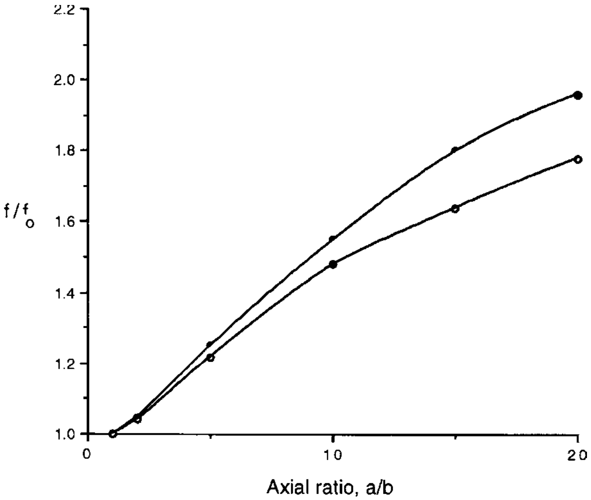


Fig. 27.1 Relationship of particle size to shape

experimental systems such as sedimentation, electrophoresis, and diffusion analysis. The interaction of the solute particle with the solvent provides information about the solute from the related measurements of *viscosity*. Viscosity is the resistance to the flow of the medium and is related to the effective volume of the solute particle as opposed to the linear dimension of the solute reflected by the frictional coefficient.

27.2.3 The Centrifuge Is a Device that Produces Motion by Generating Circular Motion with Constant Speed

A particle moves in a circular path at radius r with a constant speed. In this case, the magnitude of the velocity vector does not change but its direction does. Consequently there is constant acceleration at constant velocity. This can be shown by considering two position vectors, r_1 and r_2, at two times t_1 and t_2 and their corresponding velocity vectors, v_1 and v_2. As Fig. 27.2 shows, for very small t, the velocity change is approximately perpendicular to the velocity vectors and points to the center of the circle. In the limit, as t approaches 0, the magnitude of acceleration is given exactly by:

$$a_r = -\frac{v^2}{r} \tag{27.10}$$

The negative sign indicates that the acceleration is antiparallel to the radius vector and thus points in toward the center of the circle. a_r is called the *centripetal acceleration*. The tangential acceleration at constant velocity is 0. When the speed of a

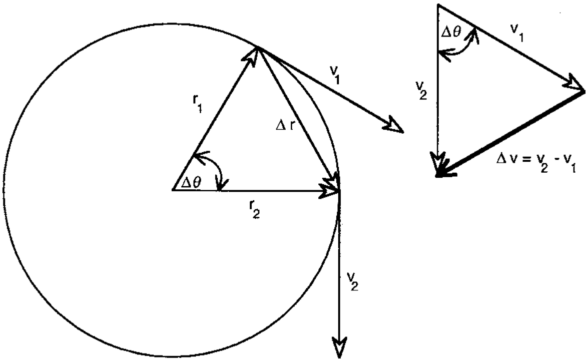


Fig. 27.2 Vector analysis of circular motion showing that the net velocity vector is directed toward the center of radius

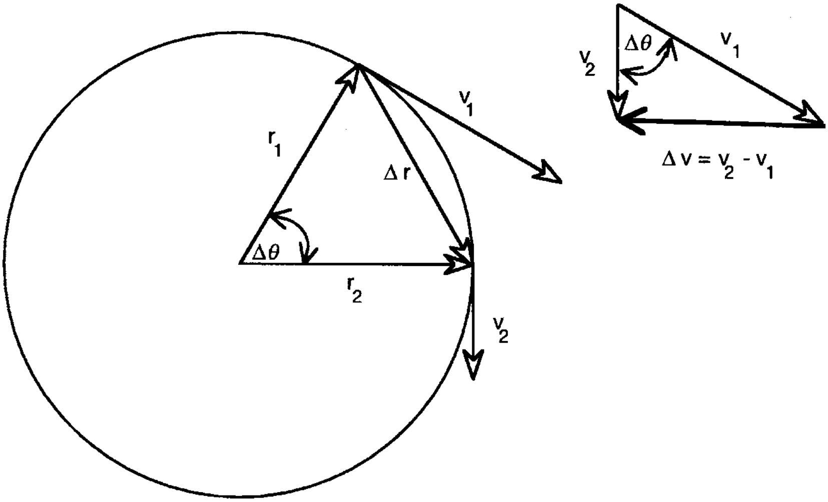


Fig. 27.3 When the velocity of a circular motion is not constant, the tangential vectors are different with respect to time and the resultant is no longer directed to the center of radius

particle in circular motion varies, the radial acceleration is still given by Eq. (27.10) but the tangential acceleration component no longer has a zero value as Fig. 27.3 indicates:

$$a_\mathrm{r} = -\frac{v^2}{r} \qquad a_\mathrm{r} = \frac{dv}{dt} \tag{27.11}$$

In cases of circular motion, the x and y motions are connected; if plane polar coordinates are used to describe the motion, the expressions of motion are often much simplified. A pair of numbers, the polar coordinates, are used to locate a point in a plane. With reference to a point of origin, any point may be described by a radial coordinate, ρ, and an angular coordinate ϕ. By convention, ϕ is measured from the $+x$-axis and increases with rotation in a counterclockwise direction. Frequently the angle is measured in radians (2π rad $= 360^{\circ}$) which are a dimensionless unit representing the ratio of arc length to radius. Motion in a circle has a fixed radius so ρ= R and the motion is described by the single variable, ϕ , which often has a time dependence, $\phi\,(t)$. The velocity is the arc length traveled in a time interval:

$$v = R\frac{d\phi}{dt} = R\omega \tag{27.12}$$

We have introduced the *angular speed*, ω , which is the rate of change of the angle ϕ. ω has dimensions of T^{-1} and is measured in radians per second in SI.

When an object is executing uniform circular motion, the time necessary for a complete revolution to occur is the *period*, T, of motion. The distance traveled in one revolution is $2\pi R$ and the period is

$$2\pi R = vT \tag{27.13}$$

and

$$T = \frac{2\pi R}{V} = \frac{2\pi R}{\omega R} = \frac{2\pi}{\omega} \tag{27.14}$$

The *frequency*, f, is the inverse of the period, $f = \frac{1}{T}$, and angular speed and frequency are related

$$\omega = 2\pi f \tag{27.15}$$

The SI unit of frequency is the hertz (Hz) which is one cycle per second.

The force experienced by a particle in a centrifuge is called the centripetal force. We have already noted that a body moving in a circle even at a constant velocity is constantly accelerating because the direction of the velocity vector is changing. We now modify the acceleration vector in Eq. (27.10) by explicitly writing the unit vector $\hat{\mathbf{r}}$:

$$a_r = -\frac{v^2}{r}\,\hat{\mathbf{r}} \tag{27.16}$$

$\hat{\mathbf{r}}$ points outward from the center of the circular motion and the acceleration vector, **a**, points toward the center of the circle. Using $\mathbf{F} = \mathrm{m}\mathbf{a}$ we can determine a force acting toward the center of the circle of rotation needed to keep the ball in uniform circular motion.

$$\mathbf{F} = -\frac{mv^2}{r}\,\hat{\mathbf{r}} \qquad 27.17)$$

This force vector is the *centripetal force* and the acceleration vector is the *centripetal acceleration*. The centripetal force on the ball swinging around your head is the tension on the string that keeps the ball moving in a circle. The magnitude of the force is given by Eq. (27.17). If the centripetal force is suddenly reduced to zero, for example, if the string breaks, the ball will no longer move in a circle but will move off in a straight line. The centrifugal force may be written in terms of the angular velocity of radians per second, ω:

$$\begin{aligned}\omega &= 2\pi f \\ &= \frac{\pi}{30}\ \text{rpm (revolutions per minute)}\end{aligned} \tag{27.18}$$

$$\mathbf{F_c} = \omega^2 rm\hat{\mathbf{r}} \tag{27.19}$$

The force is positive because it appears to act in the same direction as the unit vector. In a centrifuge, Eq. (27.19) describes the magnitude of the force acting on a particle in the centrifuge.

27.2.4 Sedimentation Occurs When Particles Experience Motion Caused by Gravitational or Equivalent Fields

We can now consider an older but still used technique, *sedimentation analysis*. In its simplest form, sedimentation analysis is practiced by setting a test tube on a countertop and allowing the force of gravity to separate the components. The red blood cells from plasma can be gently separated using a 1 g force field in this fashion. Alternatively, a centrifugal field can be created in which the particle experiences an apparently high gravitation field because it is spun rapidly in a constant circular path by a centrifuge's spinning rotor. Essentially sedimentation methods use a gravitational field (whether generated by Earth, Jupiter, or an ultracentrifuge) to make particles move and the concentration of the particles is measured as a consequence of that movement. It is much more convenient to take advantage of the non-inertial frame of reference found in a centrifuge or ultracentrifuge than to find high field gravity planets and stars; therefore most of our attention will be focused on the fictitious gravity field generated in these machines; the centrifugal force. Measurements made while the particles are moving under the influence of this field are used to determine the *sedimentation coefficient* by a process called *sedimentation velocity* analysis. The sedimentation coefficient provides information about the weight and shape of the particle. The distribution of the particles will no longer change after a certain time, and the system is said to have reached *sedimentation equilibrium*. An analysis of this set of conditions provides information about the molecular weight, density, and separate components in the system. This last point is important. Equilibrium experiments provide us with a method to separate a system into its constitutive elements which can then be further studied.

A particle in a tube that is spinning at constant angular velocity in a centrifuge will experience a centrifugal force that will push the particle away from the center of rotation. As the particle jostles its way past other molecules both of solute and of solvent there will be intermolecular interactions that retard its progress; the sum

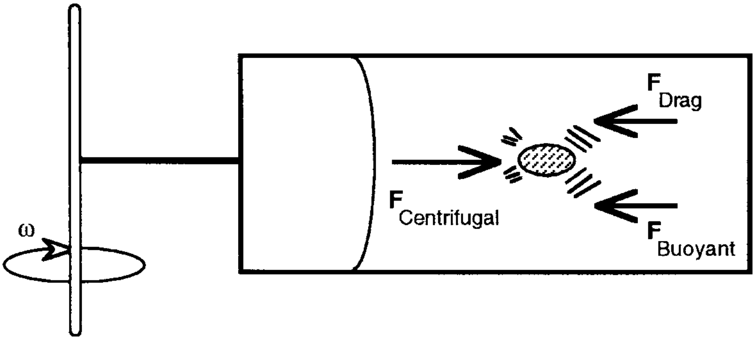


Fig. 27.4 Forces on a particle in a centrifuge

of these retarding influences is the *frictional* or *drag* force. Finally the particle will experience a buoyant force because of its differential density with respect to its solvent environment. Both the drag and the buoyant forces oppose the movement of the particle in the centrifugal field (Fig. 27.4). These nature of these three forces has been discussed in the preceding sections and we can write each of them in terms of the apparent gravitational field of the centrifuge, the centrifugal field strength, which is $g = \omega^2 r$.

$$\text{Centrifugal} \qquad \mathbf{F_c} = \omega^2 rm\hat{\mathbf{r}} \tag{27.20}$$

$$\text{Drag} \qquad \mathbf{F}_\mathrm{d} = -bv\hat{\mathbf{r}} \tag{27.21}$$

$$\text{Buoyant} \qquad \mathbf{F}_\mathrm{b} = -w_o = -\omega^2 rm_o\hat{\mathbf{r}} \tag{27.22}$$

The action of the rotor will put the molecules in motion so that a velocity is acquired in which all of the forces sum to zero:

$$\mathbf{F}_\mathrm{c} + \mathbf{F}_\mathrm{d} + \mathbf{F}_\mathrm{b} + = 0 \tag{27.23}$$

$$\left(\omega^2 rm\hat{\mathbf{r}}\right) + \left(-bv\hat{\mathbf{r}}\right) + \left(-\omega^2 rm_\mathrm{o}\hat{\mathbf{r}}\right) = 0 \tag{27.24}$$

It is customary to arrange this equation so that experimental parameters and molecular parameters can be segregated. The experimental parameters will be the angular speed of the centrifuge, the magnitude of the radius vector, and the velocity of the particle necessary to generate the drag force. The first two of these are established in setting up the centrifuge and the velocity can be measured by observing the contents of the centrifuge cell. The molecular parameters are derived from the frictional coefficient and the weight of the displaced solvent. Both of these molecular parameters depend on the size and shape of the molecule. In the case of the frictional coefficient, a broad flat or leafy structure will encounter more resistance to movement than will a compact spherical structure. The buoyant force depends on the displaced mass, m_{o}, which in turn is related to the density of the displaced fluid, the mass of the particle, and the partial specific volume of the molecule, $m\bar{v}\rho$. We proceed by rewriting Eq. (27.24) in these terms and simplifying:

$$\left(\omega^2 rm\right) + (-bv) + \left(-\omega^2 rm\bar{v}\rho\right) = 0 \tag{27.25}$$

$$\omega^2 rm\,(1 - \bar{v}\rho) - bv = 0 \tag{27.26}$$

We can immediately determine the velocity of the particle at which the net forces are zero:

$$v = \frac{\omega^2 rm\,(1 - \bar{v}\rho)}{b} \tag{27.27}$$

This equation gives a form to a series of effects that we already qualitatively know and which forms the basis of all sedimentation effects, namely

1) Particles of greater mass or higher density move faster in a centrifugal field than lighter or less dense particles.
2) The velocity of the particle varies inversely with the frictional coefficient.
3) The higher the density of the solvent, the slower the velocity of the particle.

Finally we manipulate Eq. (27.27) to achieve the sought after segregation with the molecular parameters on the left:

$$\frac{m(1 - \bar{v}\rho)}{b} = \frac{v}{\omega^2 r} = s \tag{27.28}$$

The term in the center, which is a ratio of the measured velocity of a particle to its centrifugal field, is the *sedimentation coefficient*, *s*. This coefficient *s* has units of seconds and usually the values of *s* are on the order of the 10^{-13} s. Having exponents of this order in complex equations is inevitably inconvenient and so the *Svedberg* unit is introduced. One Svedberg = 10^{-13} s and is written, S. The determination of *s* is the goal of most sedimentation velocity experiments. Brief inspection of Eq. (27.28) shows that *s* increases in proportion to the mass and inversely in proportion to the frictional coefficient and buoyancy of a particle. We anticipate that these molecular properties as measured by movement through a force field will not be constant but will vary with the concentration of the solute and the temperature of the experiment.

Consider what happens in a centrifuge tube. For simplicity first we add a homogeneous solution of buffer containing the thyroxine transport protein, transthyretin (TTR). A mixture of a material of interest, in this instance the solubilized contents of a cell, is added along with a solvent, in this case a physiological salt buffer. At rest, each particle in the tube experiences a gravitational field of 1 g while at the same time it is being buffeted by the random thermal motions of other molecules both of solute and of solvent class. Net movement is limited by the near equality of the randomizing thermal motion and the orienting gravitational field that will tend to separate the particles depending on their density and weight. When the centrifuge is turned on and reaches constant angular speed, the apparent gravitational force field experienced by the particles is clearly the dominant force. All the particles begin to move along the gravitational field lines. As the movement proceeds, the meniscus of the tube is cleared of solute and only solvent remains; however, a boundary has been formed which continues to move. The motion of this moving boundary has a rate, which is the change in radial coordinate with respect to time, $\frac{dr_{\mathrm{b}}}{dt}$. The rate of displacement is a velocity, however, and thus

$$v = \frac{dr_{\mathrm{b}}}{dt} \tag{27.29}$$

from 27.28:

$$v = \frac{dr_b}{dt} = \omega^2 r_b s \tag{27.30}$$

After integrating we obtain an expression that will allow s to be determined by graphing the measured changes in the position of the moving boundary at different time points.

$$\ln \frac{r_{\mathrm{b}}\,(t)}{r_{\mathrm{b}}\,(t_{\mathrm{o}})} = \omega^2 s\,(t - t_{\mathrm{o}}) \tag{27.31}$$

Graphing $\ln \frac{r_{\mathrm{b}}(t)}{r_{\mathrm{b}}(t_{\mathrm{o}})}$ versus $(t–t_{\mathrm{o}})$ gives $\omega^2 s$ from which s can be easily found.

The sharpness of the boundary formed by the application of the centrifugal field will be determined by the randomizing thermal forces that will tend to make the molecule move around an equilibrium position. As we discussed earlier, this amplitude of movement around the equilibrium coordinate of a molecule is dependent on a molecular parameter, the diffusion coefficient. If the tendency to be moved into action by the thermal forces is high the diffusion coefficient is large. If the molecule is massive the tendency to be moved by these forces will be very low and the diffusion coefficient will approach zero. If the diffusion coefficient is zero the boundary will remain razor sharp as it moves down the centrifuge tube. Alternatively a larger diffusion coefficient will be manifested by a boundary that broadens with the time of the experiment. Since the diffusion coefficient is a measure of the change in concentration with respect to time, a diffusion coefficient can be experimentally measured by a sedimentation method. First a boundary of high solute concentration stepping off to a (virtually) zero concentration is created which is the first event in our centrifugal field. Then after the boundary is created, the field is removed by stopping the centrifuge and the flow of the solute down its concentration gradient is followed. The diffusion coefficient, a molecular parameter, can be determined from measurements of the concentration with respect to position and with respect

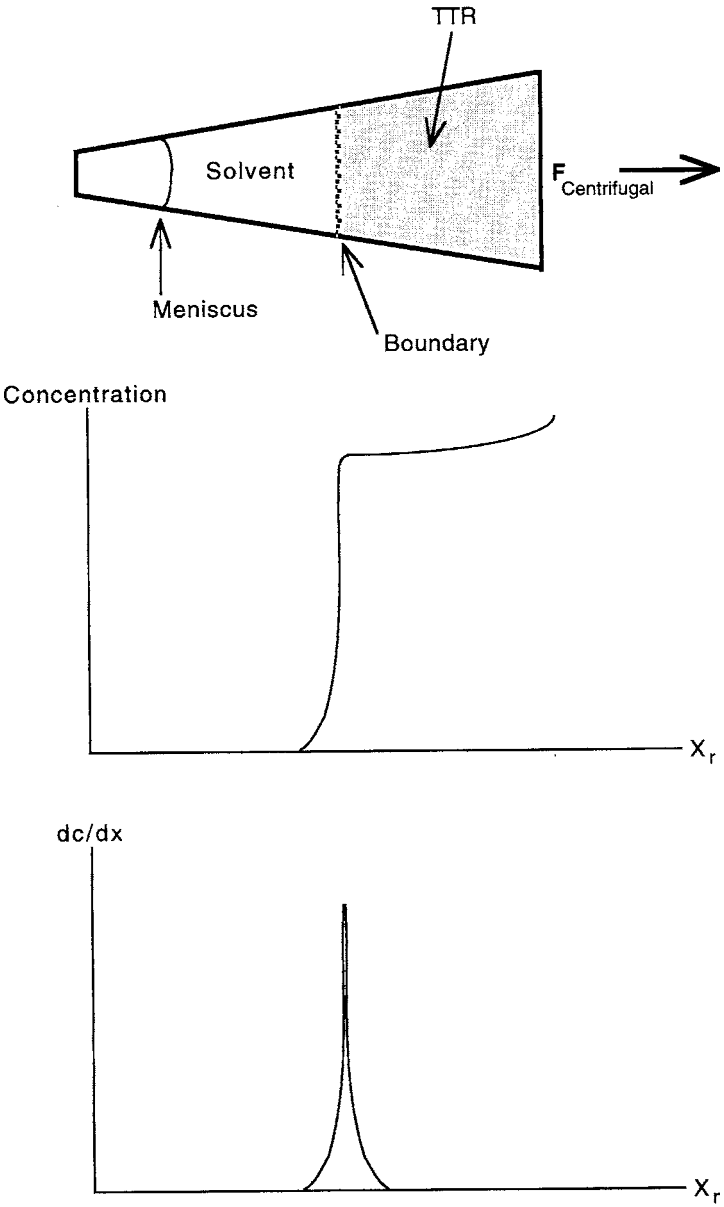


Fig. 27.5 Result of sedimentation analysis for a sample containing only tetrameric TTR

to time. This is an application of the principles of diffusion that we explored in Chapter 21.

We can now consider the results of our sedimentation experiment with TTR. As we noted earlier, biologically active transthyretin is a transport protein that carries thyroxin, the active thyroid hormone, and retinoic acid or vitamin A, in association with retinoic-binding protein. Active-TTR (TTR_4) is a tetramer of four TTR subunits that form rapidly and are normally stable in the tetrameric form. If we assume that TTR_4 is the only component of the solute in the centrifuge tube, upon application of the centrifugal field a single boundary would be formed as illustrated in Fig. 27.5. If we measure the concentration of TTR_4 with respect to the radial coordinate (dx_{r}) in the sedimentation cell, we obtain the idealized curve shown in the middle panel in which a sharp increase in measured concentration is followed by a flat plateau that remains flat until we approach the bottom of the tube where the solute is beginning to pile up. The vertical portion of the concentration curve allows us to determine *s*. The value of the relative height of the plateau will become apparent in a moment. Since we are particularly interested in the boundary between regions of different concentration, a very useful observable is the concentration gradient $\left(\frac{dc}{dx_r}\right)$ which when graphed against the radial coordinate gives the result shown in the lower panel. What happens if a second solute in present or if the tetrameric TTR dissociates into subunits during our experiment? Now we consider the case of TTR_4 and one or more potential other species. If each component in the mixture has a distinct *s* and can be separated under the conditions of the experiment, the results shown in Fig. 27.6 will be found. In the cases of the first two columns, the values of *s* can be determined for each component indicated by either the concentration curves or the peaks on the concentration gradient plots. The relative heights of the plateaus can now be used to indicate the ratio of the components. If the mixture contains a range of *s* values that cannot be separated, a broad smooth concentration curve is found indicating a heterogeneous mixture as shown in the third column.

The physical machinery used in the determination of *s* is called an *ultracentrifuge* and usually comes in one of the two varieties, *analytical* or *preparative*. The essential difference between these two classes of machines depends on whether a system is built into the centrifuge capable of detecting the changing concentration gradients during the application of the centrifugal field (i.e., when the rotor is spinning). Analytical centrifuges contain such detection systems, usually optical, while preparative centrifuges require determination of the different concentrations of solute by fractionation of the tube's contents after the experiment is completed. A schematic of an ultracentrifuge is shown in Fig. 27.7. Modern analytical centrifuges can generate effective gravitational fields of 600,000*g*. The space in which the rotor, cell, and experimental mixture are found is tightly temperature controlled usually within 0.1°C; often a vacuum is maintained during the experiment to minimize frictional forces on the rotor. The cells used in analytical centrifuges are wedged shaped and are called *sector cells*. The shape of the sector cell ensures that the lines of centrifugal force passing through the cell are parallel throughout the sample (Fig. 27.8). In

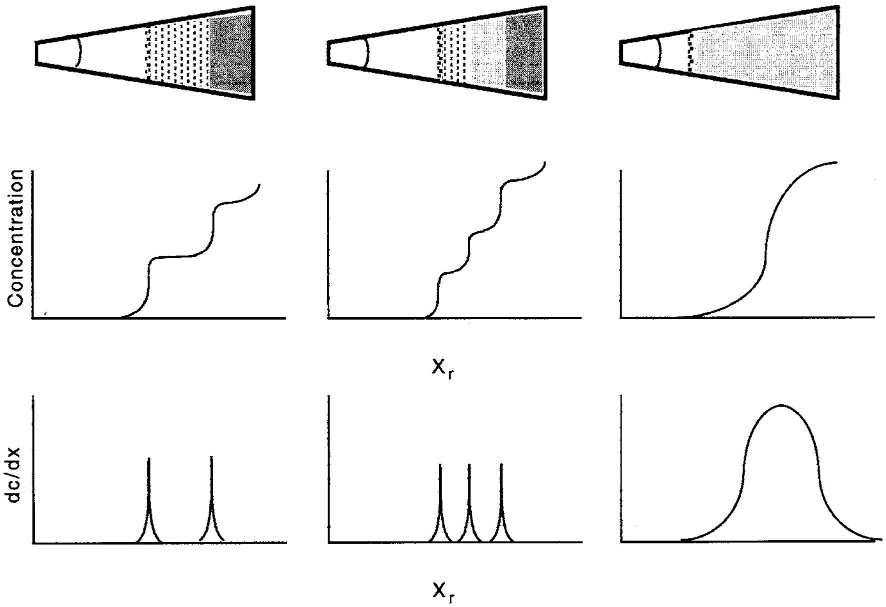


Fig. 27.6 Result of sedimentation analysis for a sample containing TTR_4 and another component

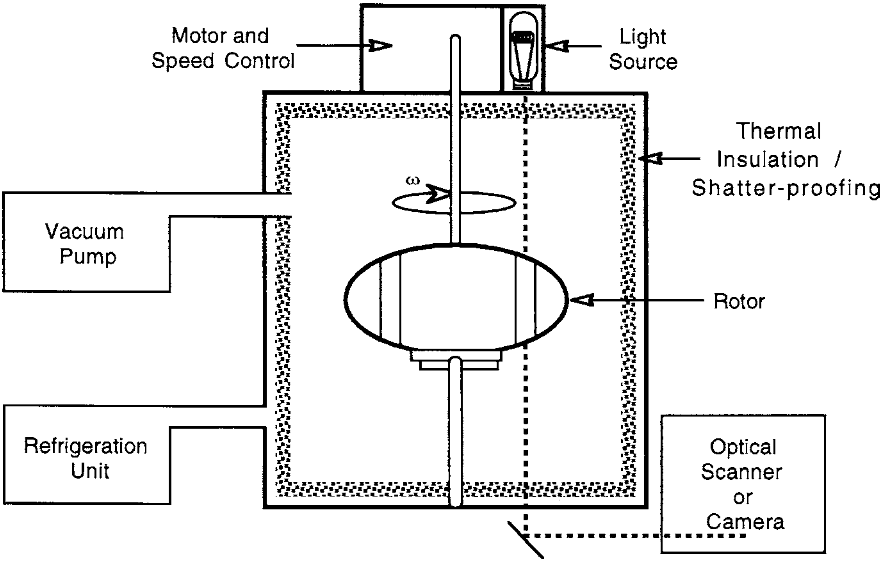


Fig. 27.7 Schematic of the ultracentrifuge

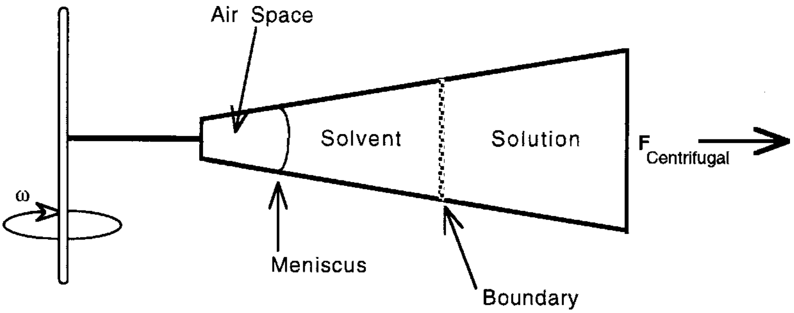


Fig. 27.8 The sector cell is designed so that the lines of gravitational force through it are parallel

addition to having the experimental chamber in this shape, the sector cell has windows on both sides so that the optical measurements for the determination of *s* can be made during the experiment. Some cells have two sectors that allow one chamber to be used as a solvent control while the solute/solvent system is simultaneously observed in the second chamber. In the preparative centrifuge, the tubes are not sector cells but are rather parallel-sided tubes. These tubes allow for more convenient collection of separated materials and sedimentation of larger quantities, but the lines of centrifugal force are not parallel and convection currents are induced in *preparative* tubes.

27.2.5 Drag Forces on Molecules in Motion Are Proportional to the Velocity of the Particle

We have discussed the physics of drag forces for particles that move through a fluid. Our discussion of forces and motion has largely treated all particles abstractly as formless and shapeless although we recognized by the assignment of phenomenological coefficients that the magnitude of the drag forces a particle experiences depends both on the shape and volume of the particle and on the viscous properties of the fluid in which motion occurs.

Drag forces act on a particle that is moving through a medium such as a gas or a liquid compared to sliding friction, which is dependent on the normal force and a constant frictional coefficient. The drag forces act in a direction opposite to the motion of the object but differ from sliding friction in that they are dependent on the velocity of the object through the medium. The experimentally determined drag force, F_D, depends, in an extremely complicated way, on the size and shape of the body. At low speeds the drag forces are linearly dependent on the speed of the particle and the *viscosity* of the medium. The viscosity, which we will discuss in some greater detail shortly, is a measure of the internal friction of the fluid medium. As the speed of the object increases, the movement of the particles causes turbulence in the medium and the drag forces increase as the square of the speed. These are the forces that dominate the drag acting on cars and skydivers. At even higher speeds, the drag depends on powers higher than v^2 but these conditions need not generally concern us. In most cases of concern to us, even in the ultracentrifuge and in high-voltage fields, the speed of molecules of biological interest can be reasonably treated by consideration of only forces linearly dependent on speed.

We have introduced the ideas of flow of solute particles into a sharp boundary of high concentration under the influence of a potential field. We also noted that the boundary would tend to become less sharp depending on the diffusion coefficient (Fig. 27.9). At equilibrium the sedimentation process will lead to a sharp boundary, which will occur when the flow of matter in this irreversible system reaches a plateau in the gradient potential field. Let us ask the question, what is the potential field in this experiment? We resort to classical thermodynamics for guidance. In a classical treatment, equilibrium in a field of force is found when the total potential is equal everywhere. This does not mean that the chemical potential is the same everywhere, but that at each point the chemical potential plus the potential energy derived from any other force field is same. In other words, in the centrifuge, we must treat the system with an eye toward both the chemical potential of the concentration gradients and the potential energy as a result of being in an apparent gravitational field, the centrifugal field. The balancing of the chemical potential field that leads to diffusion will, at equilibrium, be counterbalanced by the gravitational gradient field and net transport will stop. Similar cases in which fields other than gravity are acting can be treated in this fashion.

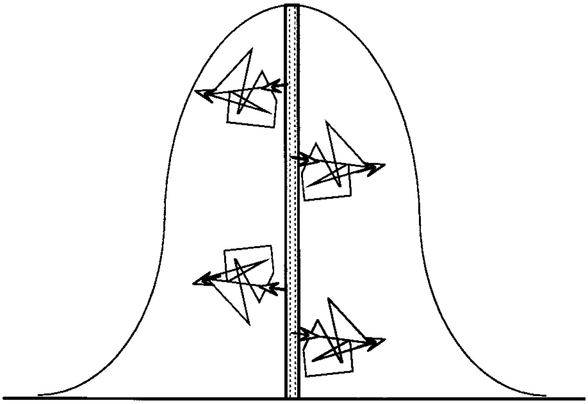


Fig. 27.9 It is the Brownian motion of molecules out of a region of high chemical potential that leads to the broadening of sharp gradients produced in centrifugation, chromatography, and electrophoresis experiments

27.2.6 *Fluids Will Move and Be Transported When Placed Under a Shearing Stress Force*

If a force is applied to a fluid, the fluid will resist the impulse to flow. This resistance is proportional to the *viscosity* of the fluid. Having knowledge of the phenomena of viscous flow is important for two reasons. First hydrodynamics is important in the mechanical flow of the predominantly liquid state of biological systems. Second, the addition of a macromolecule to a solution will alter the measured viscosity thus providing a physical method to characterize that macromolecule.

We conceptualize the flow of a liquid by imagining two very large parallel plates separated by the liquid. The upper plate is moved in the x-direction with a speed v. As the plate moves, a thin layer of fluid moves with it as a result of the friction between the plate and the layer of fluid. Imagine that the liquid is composed of many thin plates of liquid, each being moved by the frictional interaction between it and the plate of fluid above it. The layers will slide, one along the other, with a velocity gradient that is generated in the y-direction (Fig. 27.10). The liquid flowing by way of these lamina gives rise to the term *laminar flow*. This deformation of the liquid is called *shearing*. As long as the force is not too great the motion will be that of laminar flow. At greater forces turbulence will be induced and turbulent flow will result.

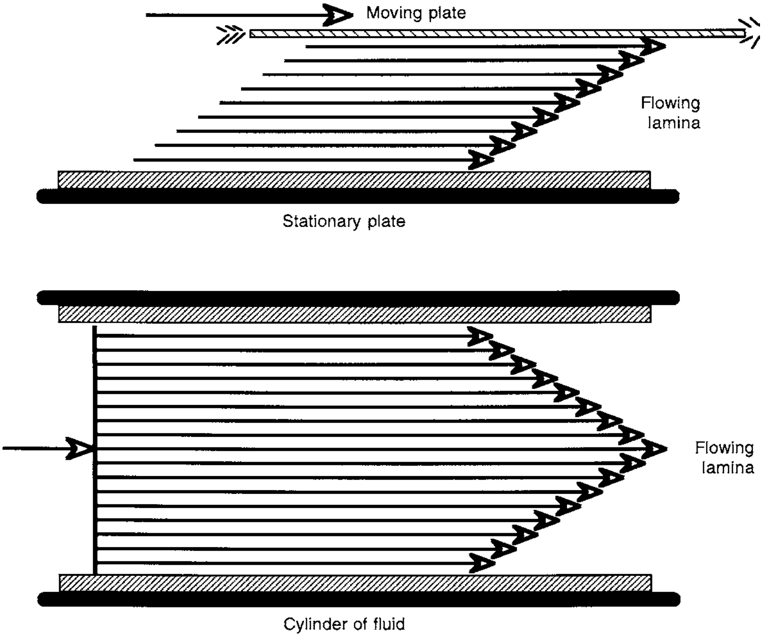


Fig. 27.10 Examples of the laminar flow of a Newtonian liquid under shearing force. *Top*, the movable/stationary plate method described in the text; *bottom*, flow in a cylinder

The frictional force between the lamina is proportional to their area and the velocity gradient generated. The coefficient that relates the gradient to the force depends on the properties of the fluid and is called the *coefficient of viscosity* (η):

$$f = \eta A \left(\frac{dv}{dy} \right) \tag{27.32}$$

This can also be written as

$$F = \eta G \tag{27.33}$$

where F is $\frac{f}{A}$, the shear stress, and G is $\left(\frac{dv}{dy}\right)$, the shear gradient or rate. Fluids in which η is constant are called Newtonian and those where η is a function of F or G are called non-Newtonian. The molecular explanation for the frictional forces at the boundaries of the capillary wall are due to the electrostatic interactions that lead to electrokinetic effects.

When a macromolecule is added to a liquid it changes the viscosity, increasing it. The change in viscosity is reflected in the relative viscosity:

$$\eta_r = \frac{\eta}{\eta_0} \tag{27.34}$$

The macromolecule may be thought of as crossing between lamina and thus increasing the friction between the planes. In 1906 Einstein showed that the relative viscosity is a function of both the size and the shape of the solute. Thus both the fraction of the solvent volume occupied by the solute and its shape enter into the determination of η.

As the concentration of a solute increases we begin to measure the solute–solute interaction and in an effort to avoid this artifact, viscosity is expressed at infinite dilution, *the intrinsic viscosity* $[\eta]$. Since the intrinsic viscosity will reflect both the molecular weight and the shape of the macromolecule, change in $[\eta]$ can be used to probe these properties of macromolecules (Fig. 27.11). The intrinsic viscosity of

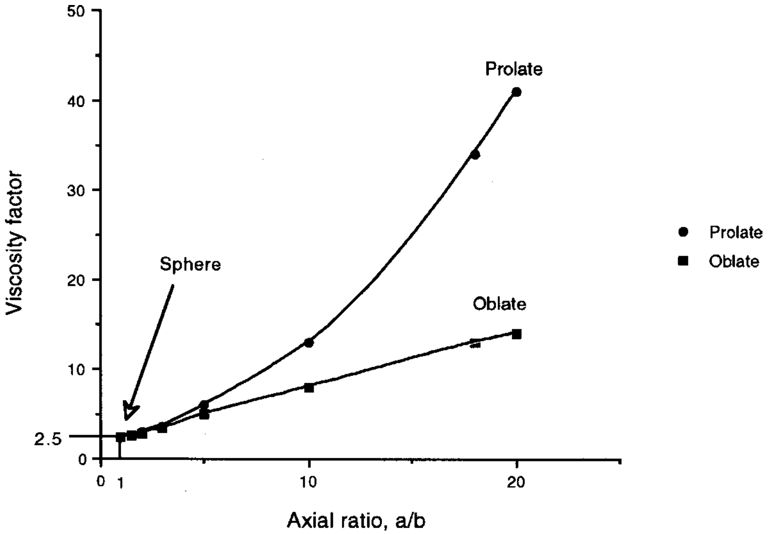


Fig. 27.11 The shapes of different molecular weight species give rise to varying viscous properties of solutions of these macromolecules. The y-axis represents a viscosity function for ellipsoids of revolution developed by Simha (1940, *J. Phys. Chem.*, **44**:25). Note that a sphere has a viscosity factor of 2.5

Table 27.2 Intrinsic viscosity of a variety of macromolecules

Protein configuration	$[\eta]$	Molecular weight
Globular		
Ribonuclease	3.4	13, 683
Albumin	3.7	67, 500
Bushy Stunt virus	3.4	10, 700, 000
Rods		
Fibrinogen	27	330, 000
Myosin	217	440, 000
Tobacco mosaic virus	36.7	39, 000, 000
Coils		
Albumin	52	67, 500
Myosin (subunits)	93	197, 000

a macromolecule in the random chain configuration is much higher than when the molecule is compactly structured (Table 27.2). Since intrinsic viscosity is essentially a measure of the volume of the solute:

$$[\eta] \propto \frac{V_{\mathrm{M}}}{M} \tag{27.35}$$

where V_{M} is the molecular volume, and M is the molecular weight. The volume of the macromolecule is related to its radius of gyration. If a θ solvent is used to dissolve the macromolecule, the intrinsic viscosity will depend on the square root of the molecular weight. Recall that a θ solvent is one in which the solution has ideal character (i.e., there are no apparent solute–solvent interactions).

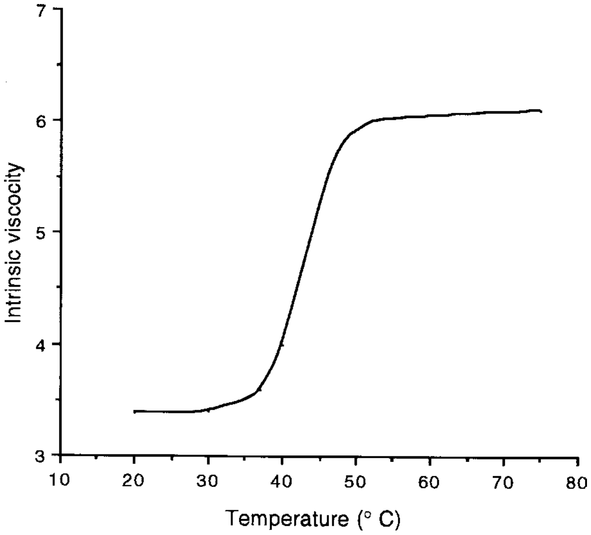


Fig. 27.12 The denaturation of a protein undergoing transition from globular to random chain configuration can be followed by measuring the change in intrinsic viscosity

A random coil configuration in a θ solvent has the character of a (pseudo)-ideal solution.

Viscosity measurements can be used to demonstrate protein unfolding because the unfolding will be reflected in an increased $[\eta]$. Figure 27.12 shows how a globular protein's denaturation to a more random chain configuration can be followed. The $[\eta]$ for a variety of different configured molecules are listed in Table 27.2.

27.3 Systems Study in the Biological Science Requires Methods of Separation and Identification to Describe the "State of a Biological System"

Identification and quantification of unknown compounds in a complex mixture are essential for numerous areas of study though none more so than the new field of systems biology, which combines the high-throughput capacity of a separation method such as chromatography or electrophoreis with mass spectrometry to give a relatively complete and complex "systems" picture of a biological system of interest. When combined with information processing methods (bioinformatics) these complex descriptions are being used to characterize the state or dynamic pattern of states, which represent the biological system under study.

The identification of the various elements of such a system can begin only following separation of its respective compounds. Once separated, compounds must be identified through their unique physical or chemical properties. Systems biology practically depends on "hyphenated" methods in which separation by methods such as electrophoresis and chromatography is followed by easily automated identification methods such as mass spectrometry. In the following sections we will explore the physical basis of these separation and identification methods.

27.4 Electrophoresis Is a Practical Application of Molecular Motion in an Electrical Field Based on Charge and Modified by Conformation and Size

We have explored some theoretical aspects of electrophoresis. Electrophoretic techniques in the modern laboratory are widely used for the separation, isolation, and identification of biomolecules. In theory we could expect the electrophoretic techniques to provide information related to structure given the relationship of size and shape to mobility but this is not the case. Practically, the complex interactions of the molecule with the ionic atmosphere of the electrolyte and the matrix through which it moves have rendered the abstractions necessary to derive the theory inadequate to provide detailed structural information. Therefore, conditions for execution of electrophoretic techniques are essentially empirically derived.

Almost all of the modern electrophoretic techniques are of the *zone* type though the technique of *moving boundary* electrophoresis was once widely used to obtain mobility data. In the moving boundary method a protein solution and an electrolyte/solvent are carefully placed together so that a boundary would form at their interface. A current is then applied that will move the boundary which can be visualized by an optical system using either schlieren or interferometric techniques. The boundary could be accurately observed and the electric field inside the apparatus could be determined from Ohm's and Kirchoff's laws. Thus by the mobility equations we have already discussed, very accurate absolute mobilities could be determined by this method. However, in spite of the accurate observable, the lack of a clear relationship to the property of interest, structure, has made the use of moving boundary electrophoresis unpopular. The moving boundary method is not well suited for separation or analysis of solutes and thus the technique has been supplanted by zonal electrophoresis techniques.

Table 27.3 Zonal electrophoresis matrix properties

Matrix	Molecules
Paper	Small molecules, carbohydrates, amino acids, nucleotides
Starch gel	Isozyme separation of proteins
Polyacrylamide gel (***PAGE***)	Wide range of proteins and nucleic acids of varying molecular weights depending on degree of cross-linking in matrix
Agarose gel	Extremely large proteins, RNA, DNA, and nucleoproteins

The principle in all zonal techniques is the use of a matrix through which a thin layer of a molecule of interest (a zone) is electrophoresed. Through the history of zone electrophoresis the support matrices have been paper, starch gel, polyacrylamide gels, and agarose gels each with properties that lend themselves to analysis of certain molecular species (Table 27.3). The advantage of zone electrophoresis is that the small amount of material that is applied to the matrix generally allows for complete separation of the solutes. As the solute moves through the matrix, the matrix serves two purposes: first it attenuates the scattering forces of diffusion and convection that will serve to blur the separation of the material, and second the matrix interacts with the solute either as an adsorbent or as a molecular sieve. The interaction of the matrix with the solute can be used to enhance the separation. The modern use of agarose and polyacrylamide gels employs the cross-linked gel as a molecular sieve which enhances the determination of the molecular weight of the solute (usually a macromolecule). The mobility of a macromolecule will depend on the net charge, size, shape, and adsorption characteristics. If an estimate of molecular weight is to be performed (Fig. 27.13) the common practice is to

1) neutralize the shape differential by denaturing the macromolecule (usually by boiling and reducing all intramolecular disulfide bonds);

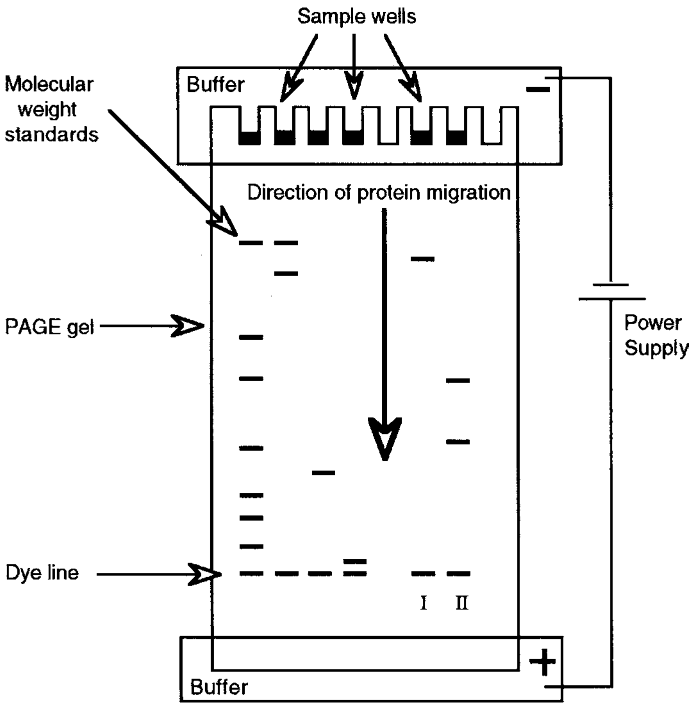


Fig. 27.13 Schematic of a gel eletrophoresis experiment. Usually a set of known molecular weight macromolecules are run in each experiment to provide a standard molecular weight scale. A fast running dye is added with the macromolecule solution to define the line of fastest migration and to keep the analyte from running off the gel. In this example the lane labeled I contains a protein that has been denatured with boiling and covered with SDS but not reduced with dithiothreitol or β-mercaptoethanol. The same protein is run in lane II except that now it has been treated with a reducing agent. With reduction of the disulfide bonds to free sulfhydryls two smaller molecular weight bands appear while the larger molecular weight band has disappeared. This result suggests that the protein is composed of two distinct smaller subunits

2) neutralize the intrinsic charge differential by covering the macromolecule with a charged amphiphile that gives it a proportional charge per unit length (Fig. 27.14). This is accomplished by mixing the denatured protein with SDS, sodium dodecyl sulfate, a detergent;
3) run the sample in an appropriately cross-linked gel to enhance the molecular sieving along with reference molecular weight samples.

The intrinsic charge of a protein can also be used to separate it from other proteins. If the protein is run in a gel with a gradient of pH, at particular pH value the net ionization of the side chains will sum to zero. This is called the *isoelectric point*. Since the movement of the protein is generated by the movement of a charged species in an electric field, when the isoelectric point is reached, the protein will stop moving. In general, after a sample undergoes separation in a pH gradient it is

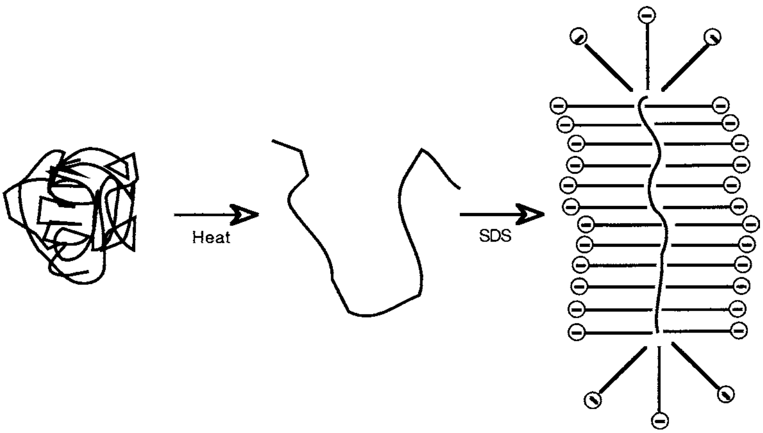


Fig. 27.14 In SDS-PAGE electrophoresis a protein is typically denatured and coated a molecule with a uniform charge per unit length by boiling in the presence of a detergent such as sodium dodecylsulfate (SDS)

re-electrophoresed under denaturing-SDS conditions to achieve a two-dimensional separation (Fig. 27.15)

In the last decade electrophoresis has become practical in very small capillaries and the practice of *capillary electrophoresis* has become an important technique for analysis of highly complex mixtures of materials. A matrix is not needed and separation occurs in the sample buffer. In many ways this technique is reminiscent of the principles of moving boundary electrophoresis. It is very fast, has high resolution, and can separate nano and picoliter quantities of sample. In addition to separating large molecules capillary electrophoresis has the capacity to separate small ionic solutes as well. Given the small size of the capillary, the interaction of the solvent with the capillary wall gives rise to an electro-osmotic effect that must be taken into account in this technique.

27.5 Chromatographic Techniques Are Based on the Differential Partitioning of Molecules Between Two Phases in Relative Motion

While electrophoretic techniques use electrical transport to separate chemical mixtures, chromatographic techniques use a flow of solvent under pressure to transport the analytes. There are a large variety of partition chromatographic techniques. All have the same goal: to separate mixtures into their specific compounds according to differences in a specific physical property typically as they move through a column or over a flat supporting matrix. The separation is achieved by allowing the analyte molecules to interact with

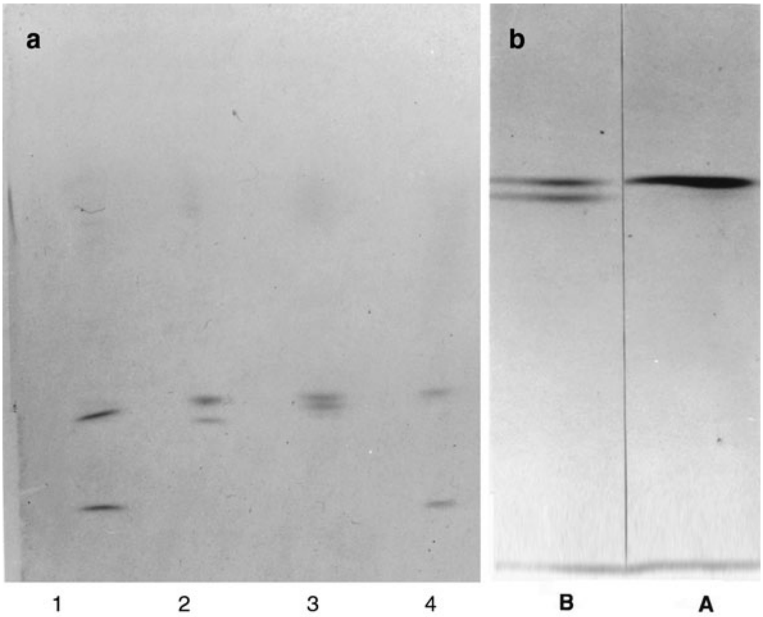


Fig. 27.15 Isoelectric focusing can separate proteins with very subtle differences. In (**a**) samples from four patients with the genetic disease, familial polyneurotic amyloidosis (FAP), are run on an isoelectric focusing gel. FAP is caused by one of a variety of single amino acid mutations in the protein transthyretin, TTR. In lane one, a histidine replaces leucine at position 58; in lane two, a methionine replaces a valine at position 30; in lane three, glycine replaces glutamic acid at position 42, and lane four is another patient with the histidine 58 mutation. The patients have both normal and abnormal protein present in their serum and both can be seen in these gels. The lower bands are the normal TTR. The isoelectric focusing gel is run after a native protein gel (no denaturants or SDS added) is first run. The TTR separated by the native gel *d* run is then cut out and electrophoresed into the isoelectric focusing gel. (**b**) The isoelectric focusing technique can be used to show the effectiveness of treatment. Here a sample of the serum from an FAP patient before (B) and after (A) treatment (liver transplantation) shows that the mutant protein is eliminated by the treatment (courtesy of Dr. Martha Skinner)

1) a “stationary phase” – compounds that remain motionless in the column typically through attachments to the column wall, and
2) a “mobile phase,” – gas or liquid phase that carries the analyte molecules over the stationary phase.

The gas/liquid of the mobile phase is also called the *eluent*. The mobile and stationary phases must not interact with each other if complete separation of analytes is desired. As the mobile phase carries the analytes down the length of the stationary phase, each type of analyte has a different degree of interaction with the molecules of the stationary phase. This degree of interaction can be represented as an equilibrium ratio, which will be called the *partition coefficient ratio*. Partitioning (or mass transfer) of analyte molecules between the mobile and the stationary phases will occur multiple times per molecule over the length of the column. Although in reality this is a continuous process, each partitioning event is treated as a separate event or *theoretical plate*. A chromatographic column is characterized by the number of these ultra-thin horizontal plates each "stacked" one-on-top of another. The difference in the degree of analyte interaction with the stationary phase results in each compound having a different time of travel past the stationary phase. The time of travel is also known as the elution or *retention time*. The relationship between retention time of a compound and partition coefficient ratio is written as

$$\begin{aligned} K_x &= \frac{[x]_{\mathrm{SP}}}{[x]_{\mathrm{MP}}} \\ &= k' \times \beta \\ &= \frac{(t_{\mathrm{r}} - t_{\mathrm{m}})}{t_{\mathrm{m}}} \times \beta \end{aligned} \tag{27.36}$$

where K_x is the partition coefficient ratio for a particular analyte x. k' is the retention factor of the analyte while β depends on the column radius and the stationary-phase plate-thickness of a particular column. k' is determined by comparing the total retention time of the analyte (t_{r}) to the eluent retention time (t_{m}). This relationship demonstrates that for a given column, the elution time for a particular analyte will depend directly on the degree of interaction with the stationary phase. The more time the analyte spends partitioned in the stationary phase, the larger the K_x and the longer the analyte retention time.

Gas chromatography, for example, utilizes a gaseous mobile phase and a high-boiling point liquid as the stationary phase. Separation is achieved by taking advantage of differences in analyte volatility (i.e., boiling points). A mixture to be analyzed must first be vaporized in a heat oven. Then, the carrier gas (typically nitrogen, helium, or hydrogen) moves the vaporized analytes to the capillary column, which is typically coated with a silicone polymer stationary phase. Following injection onto the column. The analytes then travel down the column and eventually elute based on their partition coefficient ratio. A liquid–solid partition system can be used for analytes that are more stable in liquid solution compared to a gas phase (many biological systems). The principles are the same whether liquid–solid adsorption, molecular size, ion exchange, or specific affinity are the interactions between the mobile and the stationary phases.

There are four parameters that can be altered to optimize separation in a GC system: column dimensions, stationary-phase choice, mobile-phase choice and flow rate, and column temperature. When choosing an optimal column for a particular separation, the factors to consider are the diameter of the column, stationary film thickness, and column length. Recalling that the partition coefficient ratio, K_x, depends on β, which is a function of column diameter and stationary film thickness (Eq. 27.36). More specifically,

$$\beta = \frac{r_{\mathrm{c}}}{2d_{\mathrm{f}}} \tag{27.37}$$

where r_{c} is the radius of the column and d_{f} is the stationary film thickness. If an analyte is eluting too quickly (small k'), increasing stationary film thickness would retard the analyte elution. Similarly, if an analyte is moving slowly through a given column (large k') and time is a concern for the separation parameters, retention time could be minimized by decreasing stationary-phase thickness.

When dealing with multiple analytes, the choice of column dimensions can maximize the difference between elution times. This difference is the column's *resolution*. The relationship between k' of an analyte and the resolution (R_s) between analytes can be written as

$$R_{\mathrm{s}} = \frac{N^{1/2}}{4} \frac{(\alpha - 1)}{\alpha} \frac{k'}{\left(k' + 1\right)} \tag{27.38}$$

where N is the number of theoretical plates in a column and

$$\alpha = \frac{k'_1}{k'_2} \tag{27.39}$$

When the partition constants (k') of two analytes are close together, α will approach 1, thus drastically decreasing resolution between elution times. Stationary film thickness would therefore be chosen to maximize the difference between analyte elution times. The relationship between length of the column and resolution of peaks is directly proportional and while increasing column length would also increase resolution this can be a difficult parameter to alter since as increasing column length is often costly and involves instrumentation changes.

Resolution between peaks also depends strongly on the compounds chosen as the stationary and mobile phases. Stationary-phase choice primarily consists of determining the interaction properties that will allow analytes to partition optimally: for example, a more polar stationary phase should be chosen for relatively polar analytes to allow maximal interaction and separation. Elution times and resolutions can be calculated for a particular stationary phase if analyte polarities are known. Mobile-phase choices depend to some degree on the chromatographic technique. For example, in gas chromatography, choices are limited since there are few carrier gases commercially available and certain detectors tolerate only certain gases. In general there may be more choice in the solvent composition used for a liquid chromatographic technique. Although mobile phase, stationary phase, and column dimensions are effective ways to alter resolution, another parameter that can be useful is alteration of the column temperature.

In summary chromatographic and electrophoretic techniques are useful, adaptable methods capable of separating and isolating the components of a biological sample. The resolution and time needed for separation depend on physical interactions that lead to the concentration of a chemical species into a narrow band or spatial region. The ultimate separation is affected by diffusion down this concentration gradient that begins as soon as the chemical potential of the separated species is larger than its immediate locale.

27.6 The Motion Induced by a Magnetic Interaction Is Essential for Determination of Molecular Mass in Modern Biological Investigations

The magnetic force is fundamental in the essential symmetry of our physical world and one could argue for its study solely on the basis of that beauty. However, in the twenty-first century, molecular interactions with magnetic force fields play an essential role in the determination of molecular identity using mass spectrometry and in molecular structure using nuclear magnetic resonance spectrometry. Thus exploration of electromagnetism is justified on both accounts and mass spectrometry and NMR techniques are important techniques following separation of biological samples as described in the previous sections.

Just as the early Greek *physioki* discovered that an electrical force developed when fleece and amber were rubbed together, they also knew that the mineral *lodestone* or *magnetite* could attract bits of iron. The essential rules of magnetic interactions have been known since these early observations of lodestones. We arrange them in a historical order:

585 BC Thales, the Ionian philosopher found that magnetite (found in Magnesia, now in western Turkey) would attract a certain class of materials such as iron. He concluded that magnets contained souls since they were capable of moving other objects.

100 AD The Chinese discovered that a sliver of magnetite suspended on a string would orient itself north–south.

1000 The early compass was probably finding use in navigation at least in the Far Eastern civilizations and among Arab traders.

1269 Pierre de Marincourt made a map of the magnetic field associated with a spherical magnet by probing the field with a magnetic needle. He showed that field lines surrounded the magnet and passed through to the opposite ends of the sphere. The similarity to the meridian lines of the compass with respect to the Earth were recognized.

Most experimenters with magnets had observed that all magnets have two poles and that the like poles repel one another while unlike poles attract.

1600 William Gilbert argued in his experimental treatise, *De magnete*, that the earth is a large permanent magnet with magnetic poles near the geographic poles of the Earth.

1750 John Michell used a torsion balance to demonstrate that the magnetic force varied as the inverse square of the distance between the poles. Coulomb confirmed this quantification and noted that in contrast to electrical charge, magnetic poles always occurred in pairs. Breaking a magnet led to a smaller magnet with two poles suggesting that magnetism was an intrinsic property of the magnetic material itself.

1800's Oersted discovered that flow of electric current through a wire could affect a compass needle. Ampere and others showed that wires carrying electric current could attract one another as well as iron. Ampere then proposed that all magnetism was caused by the flow of electric current either in circuits or in "molecular" current loops.

1831 Joseph Henry and Michael Faraday independently discovered that a changing magnetic field could induce a non-conservative electric current.

1865 James Clerk Maxwell unified the equations describing electrical and magnetic fields in the first of the unified field theories and wrote the equations describing electromagnetic interactions.

27.6.1 Magnetic Fields Are Vector Fields of Magnetic Force that Can Be Found Throughout Space

A magnetic vector field or **B** field is defined in the same way that we earlier defined the **E** vector field. Experimentally, a charge that is moving with velocity **v** in the vicinity of a magnet or a current carrying wire will be found to experience a force that is dependent on the magnitude and direction of the velocity of the charge. The **B** field is usually familiar, visualized by the lines of magnetic force outlined with iron filings around a magnet or electromagnet. This force, $\mathbf{F}_{\mathbf{B},}$ is associated with charges moving through space in the following ways:

1) The force is proportional to the speed of the charge.
2) The force is proportional to the magnitude of the charge.
3) The force is proportional to the magnitude of **B**.
4) The vector of the force depends on the direction of the velocity **v**.
5) If the charge moves along a line of the **B** field, there will be no force.
6) If the charge moves in any direction except parallel to the field, the force is perpendicular to both the field line and the velocity of the charge.
7) If the charge is at rest, i.e., **v** = 0, the force is zero.
8) The force on a negative charge is opposite to that on a positive charge with the same velocity.

These experimental observations are summarized in the magnetic force law:

$$\mathbf{F}_{\mathbf{B}} = \mathrm{q}\mathbf{v} \times \mathbf{B} \tag{27.40}$$

There is an angle, θ , that exists between **B** and **v** (if no angle exists, by #5, the force is zero) so the magnitude of $\mathbf{F}_{\mathbf{B}}$ can be written as

$$\mathbf{F}_{\mathbf{B}} = \mathrm{q}\mathbf{v}\mathbf{B}\,\sin\,\theta \tag{27.41}$$

with the direction of the force determined by the right-hand rule. Figure 27.16 illustrates these observations.

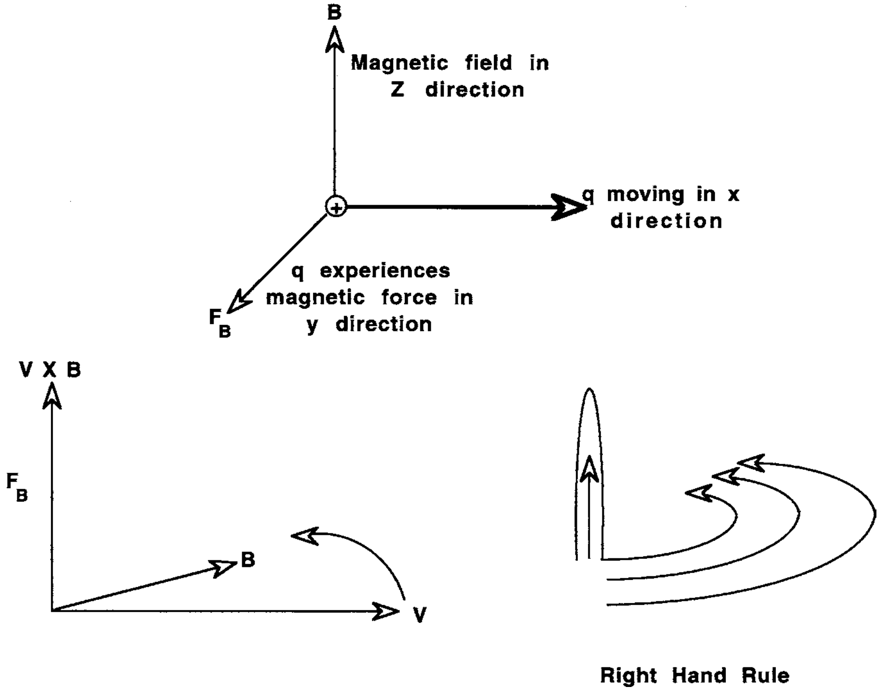

B
Magnetic field in Z direction
q moving in x direction
q experiences magnetic force in y direction
F_B
V X B
F_B
B
V
Right Hand Rule

Fig. 27.16 (*Top*) Summary of the direction of the fields and forces associated with magnetism. (*Bottom left*) The direction of the magnetic force with respect to current flow can be found by using the right-hand rule (*bottom right*) which acts to perform the operation of crossing the vector into the **B** vector

The magnetic field **B** is also called the *magnetic induction vector* or the *magnetic flux density*. The SI unit of magnetic induction is the *tesla* (T). When a charge of 1 coulomb moves with a velocity of 1 m/s perpendicular to a magnetic field of 1 T, it experiences a force of 1 N. Thus

$$1\,\mathrm{T} = 1\frac{\mathrm{N\,s}}{\mathrm{C\,m}} = 1\frac{\mathrm{N}}{\mathrm{A\,m}} \tag{27.42}$$

The unit derived from the cgs system, the gauss (G) is far more commonly used and is related to the tesla:

$$1\,\mathrm{T} = 10^4\,\mathrm{G} \tag{27.43}$$

Charges will react to electric and magnetic forces independently. When an electric field is present in addition to a magnetic field, the charge responds to the additional force, $\mathbf{F} = q\mathbf{E}$. This net force is called the *Lorenz force*:

$$\mathbf{F} = q[\mathbf{E} + (\mathbf{v} \times \mathbf{B})] \tag{27.44}$$

27.6.2 Magnets Interact with One Another Through the Magnetic Field

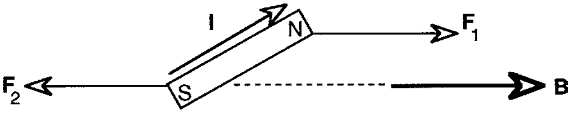
We know that if a magnet, such as a compass needle, is placed within a **B** field, the north pole of the magnet will orient itself in the direction of **B**. Figure 27.17 shows that a force is exerted on the north pole and an equal and opposite force is exerted on the south pole of the magnet. The pole strength of the magnet q^* is the ratio of the force on the pole and the magnitude of **B**:

$$q^* = \frac{F}{\mathbf{B}} \tag{27.45}$$

or in terms of a vector equation:

$$\mathbf{F} = q^*\mathbf{B} \tag{27.46}$$

Fig. 27.17 The forces exerted on a magnet generate a torque that orients the magnet to the field



There is a torque on the magnet as indicated by Fig. 27.17 which can be calculated by considering the vector **l** which points from the south to the north pole of the magnet. The torque is

$$\tau = \mathbf{l} \times \mathbf{F} \tag{27.47}$$

substituting Eq. (27.46) gives

$$\tau = \mathbf{l} \times q^*\mathbf{B} = q^*\mathbf{l} \times \mathbf{B} \tag{27.48}$$

which allows us to define the magnetic moment, μ:

$$\mu = q^*\mathbf{l} \tag{27.49}$$

and the torque on the magnet as

$$\tau = \mu \times \mathbf{B} \tag{27.50}$$

27.6.3 Current Loops in **B** *Fields Experience Torque*

When current moves in a wire a magnetic field is produced. If the wire is formed into a closed loop, by the right-hand rule, the magnetic field produced will act as a magnetic vector and the current loop will act in all measurable ways as a magnetic dipole. This dipole will experience a torque that tends to rotate the loop such that μ and **B** become aligned. The magnitude of **m** for a current loop depends only on the planar area A and the current in the loop:

$$\mu = AI \tag{27.51}$$

For a loop or coil composed, more than one turn of wire, the magnetic moment increases with each turn of the coil:

$$\mu = AIN \tag{27.52}$$

This fundamental interaction between current loops and magnetic fields underlies the physics of galvanometers and electric motors.

27.6.4 The Path of Moving Point Charges in a **B** *Field Is Altered by the Interaction*

The movement of a charge in a magnetic field is important in a wide variety of applications such as the mass spectrograph and the cyclotron. Since the force exerted by a magnetic field on a moving particle is always perpendicular to the velocity of the particle, the magnetic force does no work on the particle and does not change its kinetic energy. The magnetic force can change the direction of the velocity but not its magnitude. If the magnetic field is perpendicular to the velocity of a charged particle as shown in Fig. 27.18, the magnetic force will cause the particle to move in a circular orbit because the magnetic force causes a centripetal acceleration of the particle. The radius of the circle is given by

$$r = \frac{mv}{qB} \tag{27.53}$$

and the angular frequency is

$$\omega = \frac{v}{r} = \frac{qB}{m} \tag{27.54}$$

This frequency is independent of the radius or the velocity of the particle and is called the *cyclotron frequency*. This relationship of charge to mass is used to separate ions of different masses in the mass spectrograph (Section 27.6.5). The circular path of the charged particle occurs only when the velocity of the particle is entirely

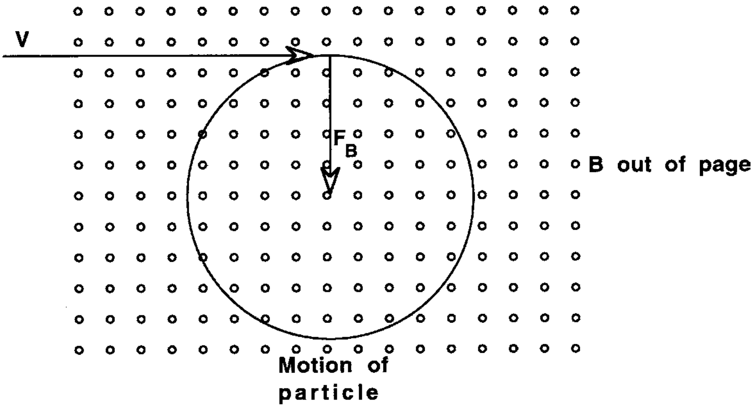


Fig. 27.18 If the magnetic field is perpendicular to the velocity of a charged particle, the magnetic force will cause the particle to move in a circular orbit because the magnetic force causes a centripetal acceleration of the particle

perpendicular to **B**. If there is any component of the velocity that is not perpendicular the particle will take a helical path.

27.6.5 The Mass Spectrometer Is Widely Used Following Various Separation Techniques to Characterize Biological Samples

Once unknown analytes are separated by chromatography or electrophoresis, a second method is needed to detect and identify each individual analyte. Mass spectrometry is designed to aid experimenters in the identification and quantification of unknown samples. A mass spectrometer separates analyte molecules based on the mass-to-charge ratio (m/z) of the gaseous ionized molecular ions. This separation results in a plot that shows the abundance of a particular molecular fragment as a function of its m/z ratio. This mass spectrum can be used on its own to determine compound identity through fragmentation patterns and molecular peak mass, or it can be compared to a database of mass spectrums made by the National Institute of Standards and Technology for thousands of known compounds. The process of mass spectrometry separation and methods of analysis are surveyed below.

The first step to separate molecules in mass spectrometry is to ionize the analyte molecules. There are multiple ways to accomplish this ionization, but two methods will be considered: electron ionization (EI), a “hard” technique, and chemical ionization (CI), a “soft” technique. In electron ionization, electrons emitted from a metal filament are accelerated toward the analyte molecules. Analyte molecules are ionized as a result of both direct collisions with electrons and by the large fluctuation in the analyte’s electric field as a result of a passing electron. Electron ionization is called a “hard” technique because it involves a high-energy transfer between the electron and the analyte molecule, which results in an increase in vibrational energy of the molecular ion. This vibration energy frequently causes the molecule to fragment into smaller, more stable portions of the analyte molecule. This fragmentation is present on the mass spectrum and is characteristic of a particular compound; often the fragmentation provides valuable clues about the structure of an unknown compound.

Chemical ionization is a second, “softer” ionization technique, meaning there is much less energy transfer associated with ionization. As a result, fragmentation of the analyte molecule is much less common with chemical ionization; it is therefore easier to see the molecular peak (M^+, the analyte 1+ ion with the same *m/z* as the analyte molecular weight) in a CI mass spectrum than an EI mass spectrum. In chemical ionization, a reagent gas (such as methane, ammonia, or isobutane) is added with the sample; this mixture is then subjected to ionization by electron bombardment in the ionization chamber. The reagent gas is present in much higher concentrations than in the sample and is therefore much more likely to interact with the electrons. Once the reagent has been ionized, it interacts both with other reagent molecules and with analyte molecules. Analyte molecules are ionized as a result of these interactions. The nature of chemical ionization leads to a greater number of hydride transfers from reagent ion to analyte ion and vice versa; it is also very common for reagent ions to react with the analyte molecule to form an ion larger than the original molecule. It is therefore very common to see CI spectrums with *m/z* values that are one higher ($(M + 1)^+$), one lower ($(M -1)^-$), or much higher ($(M + R)^+$) than the molecular peak (M^+).

Most small organic molecules (<1000 Da) can be volatilized by these type of techniques and this has lent them to study by mass spectrometry. However, historically proteins could not be studied because they were very difficult to volatilize. The recent development of electrospray ionization mass spectrometry now allows study of macromolecules including proteins. In this technique a protein is dissolved in an acidic solvent that is volatile. The solution is then sprayed into the vacuum chamber of the mass spectrometer where the solvent immediately evaporates leaving a coat of protons on the surface of the protein. The protein thus becomes a large positively charged particle that can be subjected to the separation fields.

The technique of magnetic sector mass spectrometry takes advantage of the curved motion induced in a charge that crosses a magnetic field (Fig. 27.19). In general, a sample of material under study is volatilized (vaporized) and then bombarded with electrons. As a result the molecules are broken into fragments, which are all ionized. The fragments are swept into the spectrometer where they are accelerated by an electric field and then passed into a strong magnetic field. The lightest ions will have the smallest radius of curvature. The radius increases as the mass of the ions and the strength of the electric field grow. In actual practice, the fragments are collected at a fixed sampling slit where the number of fragments is counted. The various masses are brought to the slit by sweeping the magnetic field to increasing strengths. Thus first small ions enter the slit followed by ever larger mass particles.

Electrical fields can also be used to separate the analyte once it has been ionized. There are a variety of electrostatic analyzers designed to complete this task, but we will focus on the quadrupole mass analyzer. The quadrupole consists of four electrically charged rods arranged such that there is a hole in the middle of the rods;

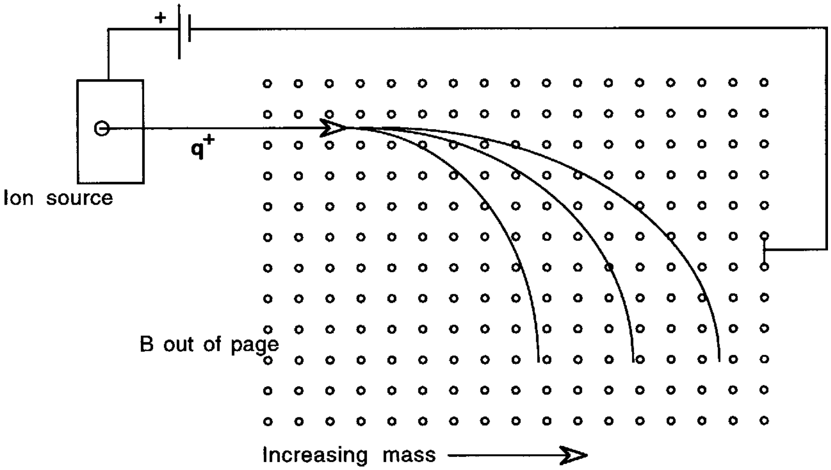


Fig. 27.19 Schematic of a mass spectrograph that uses a magnetic field to differentiate the $\frac{e}{m}$ ratio. Electrical fields (i.e., in the quadrupole trap design) can also be used to induce the motion needed to determine $\frac{e}{m}$

an ion enters at one end of the hole and the detector is at the opposite end. These rods are divided into two pairs: when one pair (A and B) is positively charged, the other (C and D) is negative. The current passed through the rods is then alternated such that this pattern of opposite charge across the pairs remains the same, but the absolute charge of each pair switches. In other words, when A and B go from positive to negative, C and D will go from negative to positive. The pairs oscillate charges continuously, thus creating a constantly changing electric field. When an ion enters the space between these rods, it will engage in a "spiraling" motion due to its attraction and repulsion to each rod. An ion can only make it down the length of the rod to the detector if the *m*/*z* is perfectly coordinated with the electric field such that when the ion is attracted to one rod, it is repelled by its opposite just enough so it does not crash into an opposing rod. For each particular *m*/*z*, there is only one frequency sufficient to ensure that the ion will make it down the length of the rods without crashing into any along the way. By altering the frequency at which the rod charges oscillate, ions with a particular *m*/*z* can be collected at different times.

There are many detectors that can monitor the *m/z* ions that pass through a quadrupole mass analyzer, but the most common is an electron multiplier dynode detector. When an ion hits the first surface of this detector, it will hit a dynode that will emit a few electrons. These electrons will travel to the next surface, which has more sensitive dynodes that emit more electrons when hit. These electrons travel to a third set of dynodes that are increasingly sensitive and release even more electrons. This method of amplifying the original response will continue for a series of 12–24 dynodes, depending on the machine. This stream of electrons creates a current that is quantified by the collector following the last dynode, which then records this current (proportional to the number of molecules present) relative to the other currents (the amount of other *m/z* ions present); this is seen as a relative abundance count on the mass spectrum.

Further Reading

General Texts

Cantor C.R. and Schimmel P.R. (1980) *Biophysical Chemistry, Part II*. W.H. Freeman, New York.

Freifelder D.M. (1982) *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*, 2nd edition. W.H. Freeman, New York.

Nölting B. (2006) *Methods in Modern Biophysics*, 2nd edition. Springer-Verlag, Berlin.

Serdyuk I.N., Zaccai N.R., and Zaccai J. (2007) *Methods in Molecular Biophysics.* Cambridge University Press, Cambridge.

Sheehan D. (2009) *Physical Biochemistry: Principles and Applications*, 2nd edition. Wiley-Blackwell, Oxford.

Van Holde K.E., Johnson W.C., and Ho P.S. (2005) *Principles of Physical Biochemistry*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

Practical Aspects of Electrophoresis

Andrews A.T. (1986) *Electrophoresis*, 2nd edition. Clarendon, Oxford, UK.

Celis J.E. and Bravo R. (eds.) (1984) *Two Dimensional Gel Electrophoresis of Proteins*, Academic, New York.

Dittman M.M., Wienand K., Bek F., and Rozing G.P. (1995) Theory and practice of capillary electrochromatography, *LC-GC*, **13**:802–810.

Laemmli U.K. (1970) Cleavage of structural proteins during the assembly of the head of the bacteriophage T4, *Nature*, **227**:680–685. (This is the original paper describing SDS-gel electrophoresis.)

Liu C., Xub X., Wang Q., and Chena J. (2007) Mathematical model for DNA separation by capillary electrophoresis in entangled polymer solutions. *J Chromatogr.*, **1142**:222–230.

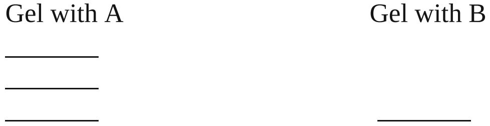
Viovy J.-L. (2000) Electrophoresis of DNA and other polyelectrolytes: physical mechanisms. *Rev. Mod. Phys.*, **72**:813–872.

Frictional Coefficient and Centrifugation

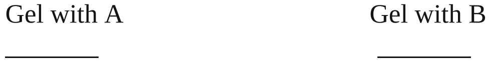
Richards J.L. (1993) Viscosity and the shapes of macromolecules, *J. Chem. Educ.*, **70**:685–689.

Problem Sets

1. Two proteins of the same apparent molecular weight by light scattering are subjected to SDS-gel electrophoresis at pH 7.5 with the following result:



Then the two proteins are subjected to SDS-gel electrophoresis at pH 9.8 with the result:



What can be deduced about the composition of these proteins? What do you think protein A could be?

2. The difference between normal hemoglobin A and the sickle-cell hemoglobin mutant protein is a single amino acid replacement of glutamate with valine in the β-chains. (Each hemoglobin molecule is composed of 2α- and 2β-chains). The mobility of these two proteins can be measured as $+0.3 \times 10^{-9}$ and -0.2×10^{-9} m^2/s V.

   (a) Match the protein with its mobility.
   (b) On an isoelectric focusing gel which protein will be found at the more acidic position in the pH gradient?

3. On average, the addition of SDS $[H_3C–(CH_2)_{10}–CH_2O–SO_3{}^{-}\ Na^+]$ to a denatured protein leads to a stoichiometry of 2 amino acids/1 SDS molecule. Estimate the net charge on the following proteins following denaturation. (A reasonable rule of thumb is to assign each amino acid a weight of 100 Da.)

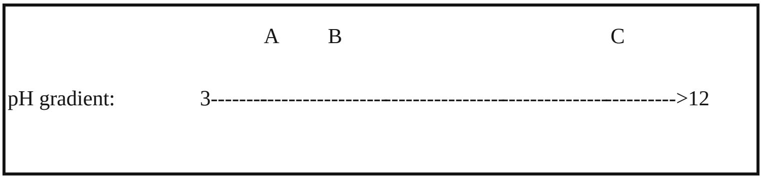
   (a) Albumin (67.5 kDa)
   (b) Ribonuclease A (12.4 kDa)
   (c) Fibrinogen (330 kDa)
   (d) Hemoglobin (68 kDa)
   (e) Myoglobin (18 kDa)

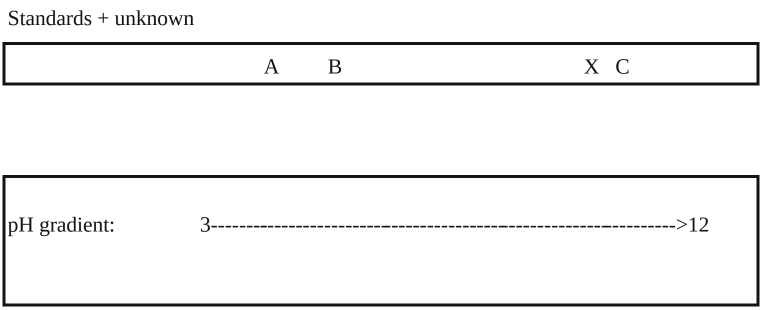
4. You have isolated a new protein (X) from cellular fragments. You run an isoelectric focusing gel to help characterize the protein. Three standards (A, B, C) are run along with the new protein. The standard's pI_s are given below.

Standards name	p*I*
β-Lactoglobulin	5.2
Cytochrome *c*	10.6
Albumin	4.8

After running your gel you realize that you have lost the key to your standards and so run a separate gel to figure out their order. The two gels (one standards only and one standards + unknown) have the following order:

Standards:





(a) Make a new key for the standards.
(b) What conclusions about your unknown protein can you make from this experiment?
(c) What further experiments could you now perform to further characterize the physical nature of the protein?

5. Proteins with subunit chains linked by cysteine–cysteine disulfide bonds (RS–SR) can be reduced to sulfhydryl (R–SH··HS–R) containing unlinked chains by the addition of β-mercaptoethanol or dithiothreitol. These reduced proteins will then run independently according to their molecular weights on an SDS denaturing gel. The following pattern is obtained for a mixture of proteins:

	Non-reducing lane	**Reducing lane**
a	-------------	-------------
b	-------------	
c	-------------	-------------
		-------------
d	-------------	-------------

Running buffer	-------------	-------------

(a) Which of the proteins in the original sample are linked by disulfide bonds?
(b) If the molecular weight of (d) is 18,000 Da and (a) is 130,000 Da what is the approximate weight of the unknown protein and its fragments.
(c) How many subchains make up the whole protein? Who could you support your argument?

Chapter 28
Analysis of Molecular Structure with Electronic Spectroscopy

Contents

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_28, © Springer Science+Business Media, LLC 2010

28.1 The Interaction of Light with Matter Allows Investigation of Biochemical Properties

We have constructed a view of biochemical state space based predominantly on the ideas of a limited set of mechanical motions and the existence of electrical charge. Using just these basic concepts we have been able to develop useful descriptions of light and electromagnetic radiation and subsequently of atomic and molecular structure. These ideas are developed from the same core physical concepts. Therefore it is not surprising that a substantial interaction between electromagnetic radiation and matter will be found. A great deal of structural information can be gained by studying these interactions. In general we can explore structure of biological molecules by either focusing on either the quantum mechanical aspects or the wave-like nature of light. The first forms the basis of spectroscopic experimental methods and is the subject of this chapter and the second, the basis of light scattering (e.g., x-ray diffraction) and is the subject of Chapter 29.

28.2 The Motion of a Dipole Radiator Generates Electromagnetic Radiation

We discussed electrical dipoles in Chapter 7. The connection between electrostatics and electromagnetics is seen when a dipole is put into motion. A dipole is a separation of charge in a linear dimension and we can imagine that, when the dipole charge is motionless, associated electric field lines extend into space. If the charge is moved up and down, the charge lines will also move in a time-dependent fashion. If the electric charge oscillates with a harmonic function, the dipole moment and its associated lines of electric force will generate an electric harmonic wave. The changing electrical field itself induces a magnetic field that is propagated normal to the electrical wave. The changes in the electric and magnetic fields do not happen everywhere instantaneously but rather are propagated outward from the dipole at the speed of light. The frequency and the wavelength of the electromagnetic wave are determined by the characteristics of the dipole radiator. For electromagnetic radiation such as visible light, ultraviolet light, X-rays, and γ-rays, the radiators are atoms and nuclei and the radiation is dominated by quantum mechanical rules. For longer wavelength radiation, the dipole radiator may be considered as having the dimensions of an antenna and the propagation can be treated by the wave abstraction we described in some detail in Chapter 7 and will explore in greater depth in coming sections.

28.3 Optical Interactions Can Be Treated at Varying Levels of Abstraction

The field of optics is generally broken into three areas: geometrical, physical, and quantum optics. Each of these fields serves the biophysical researcher in different situations and familiarity with both the applications and limitations of each is valuable. Geometrical optics essentially treats light as light rays while physical optics is essentially concerned with the nature of light and treats light as waves. Quantum optics treats light as a photonic particle with properties of both a wave and particle and is concerned with the atomic interactions of light and photons. Geometric optics applies in situations in which light can be considered on a macroscopic scale. This means that we represent light as a ray that (1) travels in a straight line and (2) is reflected and refracted at the boundary between two transparent substances. Physical optics is needed conceptually when dealing with the microscopic nature of light and in fact the classical wave-like nature of light and its interactions. The boundary between geometric optics and physical optics may best be appreciated by conducting an experiment in which we attempt to isolate a single light ray. Figure 28.1 illustrates this case.

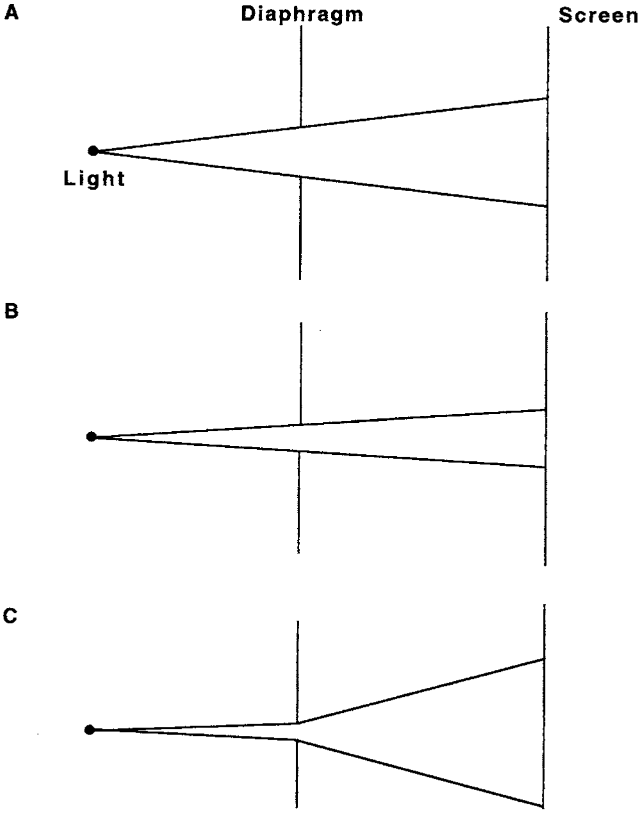


Fig. 28.1 Experiment to explore the difference between geometric and physical optics

We use a bright single point source of light (L) such as an arc lamp and by using small diaphragms (D), attempt to focus a single ray of light on the screen (S). Initially, when the hole is reasonably large, a bright spot appears on the screen whose size and shape can be predicted from the geometrical construction of drawing the light from the source as rays that travel in straight lines. Only the rays that pass directly through the diaphragm will strike the screen. Now the diaphragm is narrowed and the bright spot on the screen narrows proportionally. So far so good – we are on the trail of a single light ray. Finally, the diaphragm is narrowed still more to a dimension of several hundred microns but instead of isolating a single ray of light, the physical dimension of the spot grows larger although with a feeble light! We have reached the limit of geometrical optics and have demonstrated diffraction of the light. If we look back at our earlier experiments now, we also notice that the edges of the spots of light on the screen were never quite sharp, these edges are also the result of diffraction. So in fact diffraction was occurring all along but at the larger diaphragm sizes the ray-like qualities of light dominated and we ignored the diffraction effects. Diffraction (see earlier discussion) is a consequence of the wave nature of light and thus is the domain of physical optics. In many biological applications we cannot ignore diffraction and so must have a good working knowledge of physical optics but there are also many cases in which we can ignore diffraction effects and thus use the techniques (in some sense, the parlor tricks) of geometrical optics.

This splitting of optics into three “fields” is an excellent example of the relationship between the real state space, formal models of varying degrees of abstraction, and the empirically validated observable. Obviously all three of these fields are looking at the same state space and the most accurate/complete model is the wave–particle duality system. However, when the wavelengths are very small compared to the dimensions of the equipment available for study and the photon energies are small compared with the energy sensing capacity of the instrumentation, geometrical optics is a useful though crude approximation of electromagnetic behavior. If the wavelengths are of comparable dimension to the equipment, however, but the photon energies remain at or below detectable limits, then this is the domain of classical electromagnetic theory or physical optics. When at very short wavelengths the wave character essentially disappears and if the equipment is sensitive enough that the photon energies are large in comparison to those thresholds, then the relatively simple photon picture is useful. We will now take up the quantum level of interaction. Here light considered at the photon level interacts with molecular structure.

28.4 Atomic and Molecular Energy levels Are a Quantum Phenomenon That Provide a Window on Molecular Structure

Figure 28.2 shows an example of a molecular spectrum. Inspection of these tracings shows that three features stand out:

1) There are points of maximum inflection occurring at particular wavelengths.
2) Each maximum has a different intensity
3) The maxima are spread to some degree and are not sharp.

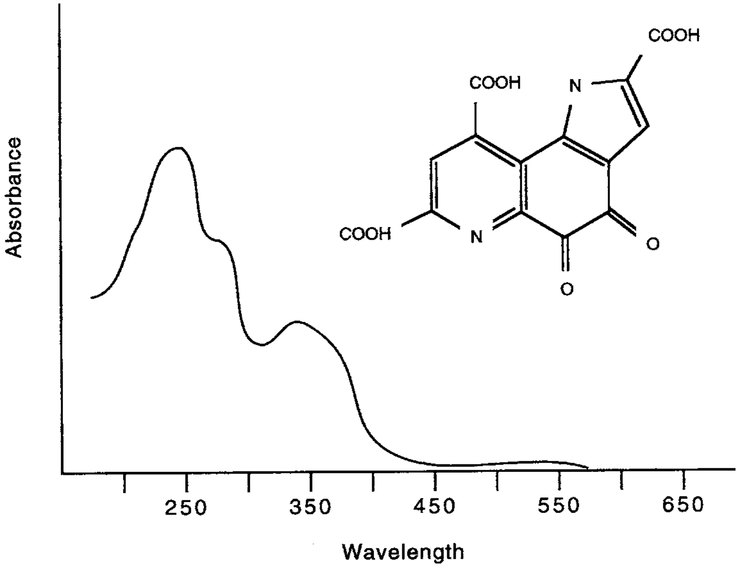


Fig. 28.2 Molecular spectra of pyrroloquinoline quinone, a redox cofactor in a variety of bacterial enzymes

These features are connected to the fundamental phenomena underlying the interaction of light and matter and are also the practical handle that we use daily in the research laboratory in our exploration of biochemical structure and function. We will explore each of these features to establish the general principles for all types of spectroscopy and will then briefly examine how these same features can be used to provide practical information about biochemical state space in daily practice.

28.4.1 There Are Points of Maximum Inflection Occurring at Particular Wavelengths

We learned earlier that the absorption and emission of radiation requires a quantum mechanical treatment to account for the observed spectral lines and the phenomena of the photoelectric effect, fluorescence, phosphorescence, and lasers. The energy in a photon that will be absorbed or emitted is given by Planck's law:

$$E = hv \tag{28.1}$$

The energy levels in a molecular system are discrete and quantized. Absorption of radiation can only occur at wavelengths corresponding to the discrete energy separating these levels. In Chapter 8 we explored some of the arrangements of these levels for the different mechanical motions of molecules (i.e., translation, vibration, and rotation). The absorption (or emission) peaks seen in spectroscopic studies are related to these discrete transitions. Yet it is clear that not all of the possible transitions have spectral lines associated with them. It is useful to think of all the energy jumps as discrete levels to be populated and to assume that the intensity of a given line will depend on its relative rate of population. Thus the intensity of a line is determined by the likelihood of a transition occurring in a particular amount of time (i.e., it is a rate-dependent phenomena).

Before we examine what can be learned from spectroscopy in biological systems, consider the energy level arrangements of a simple diatomic molecule. Figure 28.3 shows the potential energy curve for such a molecule with respect to the internuclear separation of the nuclei in both ground and excited states. The energy of the electronic state is given by this potential energy curve. In addition to the electronic energy states for this system, a diatomic molecule has vibrational and rotational energy modes. The vibrational levels depend on the shape of the electronic energy curve and are more widely spaced than the rotational energy states. Transitions are permitted between these energy levels, and the rates of the transitions are determined by the quantum mechanical selection rules. The energy associated with a

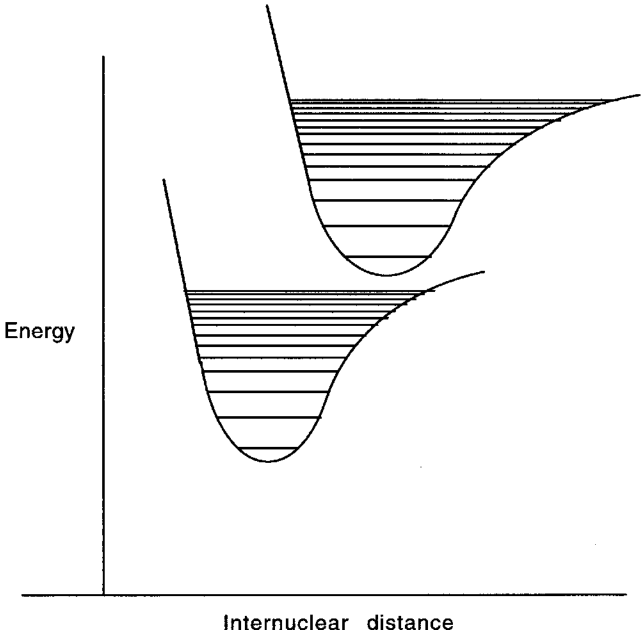


Fig. 28.3 Energy curve and levels for a prototypical diatomic molecule. A ground state and first excited state of an electronic orbital is shown with the *horizontal lines* representing the energy levels of the accessible vibrational energy states. Rotational energy states are not shown but lie between the vibrational levels and are much more closely spaced

given transition is associated with a specific frequency photon as given by Planck's law. The different energies associated with the wavelength bands of the electromagnetic spectrum are associated with different motions in the molecules (i.e., the frequency associated with an electronic transition falls in the UV–visible range while the bond vibrations are associated with energies that fall in the infrared band) (Fig. 28.4). In general, rotational spectra are useful only in the study of molecules in the gaseous state, and so microwave spectroscopy has little role in biochemical investigations. The likelihood of a photon being absorbed depends on the probability of a transition occurring which is a phenomenon governed by quantum kinetics. Just as the absorption of a photon is associated with transitions to higher energy level and is the basis of absorption spectroscopy, the emission of a photon allows movement to a lower energy level and accounts for the phenomena of fluorescence and phosphorescence.

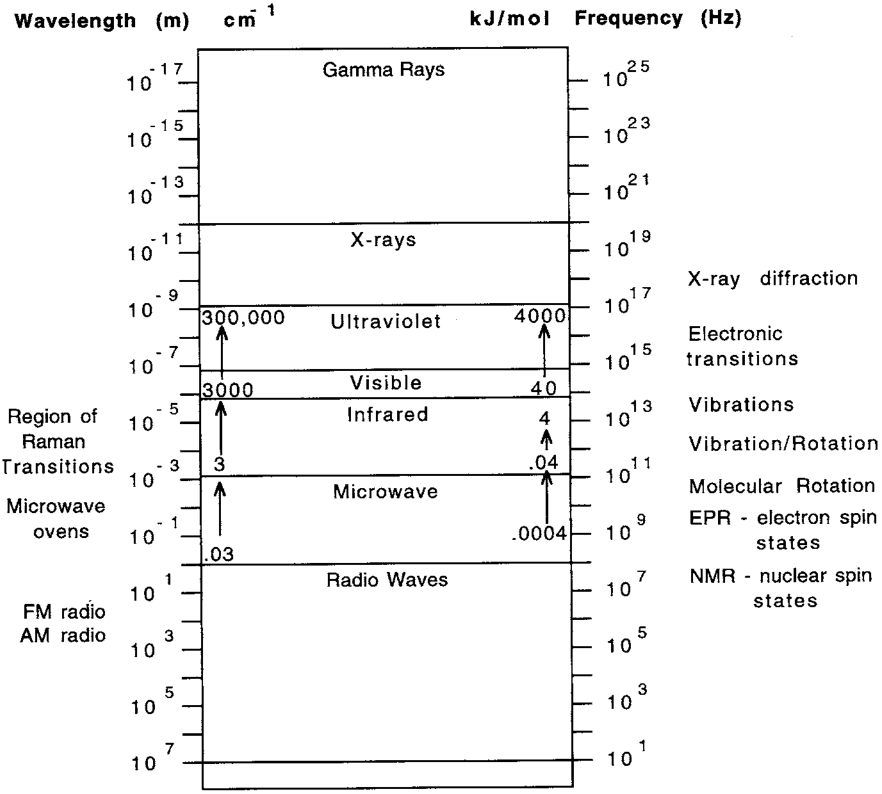


Fig. 28.4 Bonds and transitions in biological systems. The electromagnetic spectrum and associated atomic/molecular motions at varying transitions. The spectra are described in terms of wavelength, frequency, and wave number $v' = 1/\lambda$

28.4.2 Each Maximum Has a Different Intensity

In our discussion of quantum mechanics we explored some aspects of atomic spectra which are due to electronic transitions from the ground state, *g*, to the excited states, *e*. These transitions depend on the interaction of the electromagnetic field with the electronic structure of the atom. A reasonably simple sketch of this process is as follows. The electronic distribution for a given energy level can be derived from the Schrödinger wave function. Different energy states have different distributions and different shapes. Recall that the shapes of the *s*, *p*, and *d* orbitals are physically interpreted as representing different charge distributions, all asymmetric except for the *s* orbital. Each orbital has an associated dipole moment, zero in the case of the *s* orbital, and finite dipole moments with *x*, *y*, and *z* coordinate character in the case of the *p* orbitals. Like the electric distributions discussed in Chapter 7, these electronic orbital distributions are able to interact with another electric field. These interactions perturb both the electronic distribution and the electric field. The coupling of the electric perturbation to the electronic distribution occurs in such fashion that the shape (more precisely the symmetry) of the distortion is picked up by the system. If the interacting electric field is time independent (i.e. static), the perturbation will cause a widened separation of the energy levels when the ground state is compared to the excited state. When there is a wide energy separation between two states, the perturbation has little effect on increasing the separation. Alternatively, a narrow energy separation in the ground state responds to a perturbation with a wide separation in the excited state. A similar effect on the energy levels of an atom or molecule can be seen in the presence of a magnetic field.

Before expanding our discussion to spectroscopy in which the electromagnetic perturbation is time dependent because of the wave nature of light, we must consider the question: How do we treat the transition of an electron as it moves between two energy states given by the orbital quantum numbers *n*, *l*, and m_l? This is important because all of the potential transitions from one energy state to another (such as predicted by the Balmer, Lyman, and Paschen series) are not found to occur and even those that do occur do so with varying intensity. This suggests that there is a selection of certain transitions while others may be forbidden. These *selection rules* determine the observed spectrum of an atom or molecule. In the case of electronic transitions, when the electronic charge distribution between the ground and excited state changes, the dipole moment of the molecule also changes. We can consider the transition state as a combination of the ground and excited states' shape and distribution. This transition electronic state has an associated *transition dipole moment*. The transition dipole is the coupling mechanism between the molecule and the perturbing electrical vector. If the electron distribution changes symmetrically between the two states, the transition dipole will be zero, there will be no coupling of the perturbing field with the molecule, and the transition will be forbidden. In quantum mechanical terms this means that it will occur at a very low rate usually about 1000-fold less-often than the allowed transitions. This is why $s \rightarrow s$ transitions are not seen. Alternatively, if the change in state involves an asymmetric electronic redistribution, the transition is allowed and a spectral line is found. An example of such a change is an $s \rightarrow p_z$ transition. Although these examples are for atomic orbitals, similar arguments apply to molecular orbitals and only asymmetric molecular transition dipole moments give rise to molecular spectra. The transition dipole moment acts as the coupling antenna, whether a photon is absorbed from or emitted to the radiation field.

The adsorption and emission of photons of energy necessary for rotational and vibrational energy transitions can be treated in a similar fashion. In these cases a molecule must possess a permanent dipole that is capable of creating or interacting with a varying electrical field when it rotates or when it vibrates. A vibrational motion that does not change the transition dipole moment is a forbidden transition.

The selection rules derive from the conformance of the photon–molecular interaction to the laws of conservation, invariance, and symmetry. A *gross selection rule* relates to the overall character that a molecule must have in order to interact with the electromagnetic field and undergo a particular type of transition. We have already implied the gross selection rules for electric dipole-induced transitions which require the following:

1) for *rotational* transitions: that molecules possess a permanent dipole moment that can be rotationally accelerated because of an applied torque;
2) for *vibrational* transitions: that the molecule has a changing dipole moment that can be coupled to, can be excited by, or can generate an electromagnetic radiation field;
3) for *electronic* transitions: that the transient dipole moment is non-zero.

Specific selection rules define the transitions that are allowed between different quantum numbers characterizing a system. Usually the specific selection rules are related to conservation of angular momentum and symmetry requirements. For example, a photon possesses spin ($s = \pm 1$) and thus is either left- or right-hand polarized. When a particular photon interacts with an electronic distribution, it imparts its angular momentum to the electronic structure. Likewise, when a photon is emitted from a molecule it takes angular momentum with it. This spin must be conserved after the interaction, and hence the $s \rightarrow d$ transition of an electron is forbidden.

28.4.3 The Maxima Are Spread to Some Degree and Are Not Sharp

The energy levels available for a given electronic shift include lines for each rotational and vibrational mode similar to those indicated in Fig. 28.3. If each of these transitions were made distinctly, the spectra around each maxima would be exceedingly complex. However, most biomolecules are studied in solution, and the repeated collisions with the solvent lead to a complete blurring of all of the rotational transitions. If molecules were studied in hydrocarbon solvents many of the vibrational transitions would be apparent, but in interacting solvents such as water these structural details are lost. The transition probabilities for an electron going from the lowest vibrational state in a ground-state orbital to a vibrational level in the excited state are treated by using the Franck–Condon principle. This principle asserts that the electronic transition is so much faster than the nuclear rearrangement going from the ground to the excited state that the nuclear separation (and its potential energy curve) can be treated as unchanging. An electron will make a vertical transition and will prefer to go into a vibrational mode that is most similar to the one from which it is being excited (Fig. 28.5). The various intensities for each of these vibrational transitions will depend on how much alike are the excited and

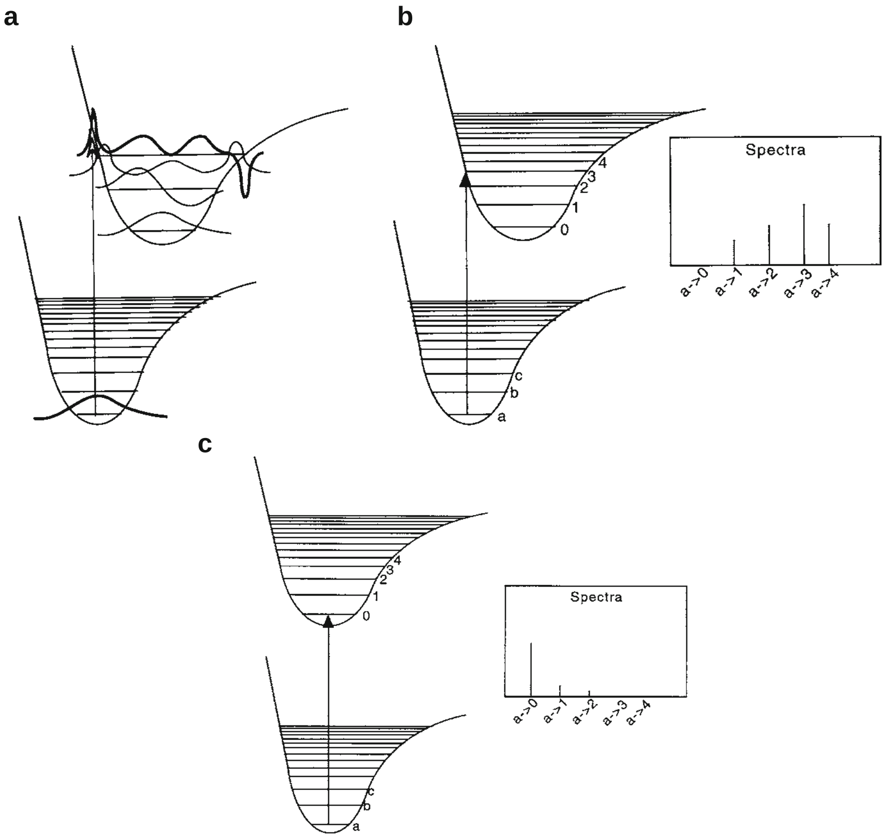


Fig. 28.5 **(a)** The Franck–Condon principle like the Born–Oppenheimer approximation assumes that the speed of electron motion is so much faster than nuclear movement that internuclear motion can be ignored during electronic transitions. The likelihood of a transition depends on how closely matched the internuclear distances are between the ground state and the excited state. (**b**) The probability of a transition and hence the intensity of the spectral line depends on the likelihood of finding the system at a particular internuclear distance. Compare the spectra and the intranuclear distance of B1 and B2

ground states (Fig. 28.5b). The various rotational and vibrational states are smeared by the solvent interactions, and this condition leads to the broadening of the spectral maxima.

Other phenomena will lead to broadening of spectral lines and are mentioned here for completeness. If the molecule making the transition is free to move to or from the observer at sufficient speed, a *Doppler shift* of the apparent frequency will be seen with broadening of the spectral line. This effect is generally seen in gases and probably has little importance in aqueous solution at room temperatures. However, even at temperatures too low for significant Doppler broadening, a spectral line does not have infinite sharpness. This phenomenon is due to the indeterminacy principle that acts through the mechanism of *lifetime broadening*. Lifetime broadening occurs because when the Schrödinger equation is solved for a system that varies with time, the exact energy changes cannot be precisely specified. If an excited state has a finite lifetime, τ, the energy of the associated transition can only be known to a precision of the order of $\delta E \approx {}^{h}/_{\tau}$. Short-lived excited states will therefore have broad spectral lines, while long-lived excited states will have much sharper spectral lines. Events that act to shorten the lifetime of an excited state will lead to broadening of the lines. The two principal mechanisms that shorten these lifetimes are stimulated emission, which is the mechanism of laser action, and spontaneous emission, which is “simply” the process by which an excited state discards energy and returns to the ground state.

We will see in our upcoming discussion of fluorescence, phosphorescence (and have already seen in our explorations of charge transfer) that excited electrons can discard energy in several ways through discrete energy exchange. However, one of the most common processes that can remove energy from the excited state is collisional deactivation, which occurs when the energy is transferred from the excited state to molecules (often of solvent) that collide with the excited molecule. The collision of the excited state with solvent or with other molecules in the environment will shorten the lifetime of the excited state and lead to a broadened spectral line. Finally, we must linger here for a moment because the idea of spontaneous emission is most fundamentally explained through quantum electrodynamics. Quantum electrodynamics treats the electromagnetic field as if it were a quantized assortment of harmonic oscillators. The quantum mechanical treatment of a harmonic oscillator leads to the conclusion that the oscillator is never at rest because it always possesses a zero-point energy and hence even a vacuum possesses a granular electromagnetic energy field. An excited electron, being charged, will interact with these grainy regions of field energy and “spontaneously” emit a photon.

28.5 Absorption Spectroscopy Has Important Applications to Biochemical Analysis

In summary, the maxima of absorption and emission are due to the discrete energy jumps between atomic and molecular orbitals. The varying intensities depend on the rates of transition determined by the selection rules, and the broadening of the

Table 28.1 Chromophores of importance in biological systems

Structure	Transition	Absorption maximum	Extinction coefficient (nm) (ε)
–COO–R	$n \rightarrow \pi^*$	205	50
	$\pi \rightarrow \pi^*$	165	4.0×10^3
C = O	$n \rightarrow \pi^*$	280	20
	$n \rightarrow \sigma^*$	190	2.0×10^3
	$\pi \rightarrow \pi^*$	150	
C = S	$n \rightarrow \pi^*$	500	10
	$\pi \rightarrow \pi^*$	240	9.0×10^3
–S–S–	$n \rightarrow \sigma^*$	250–330	1.0×10^3
C = C	$\pi \rightarrow \pi^*$	175	8.0×10^3
Pyrimidine	$n \rightarrow \pi^*$	300	325
	$\pi \rightarrow \pi^*$	245	2.0×10^3
Purine	$\pi \rightarrow \pi^*$	220	3.0×10^3
	$\pi \rightarrow \pi^*$	265	8.0×10^3
Name		Absorption maximum	Extinction coefficient (nm) (ε)
Nucleic acids			
	Adenine	260.5	13.4×10^3
	Adenosine	260	14.9×10^3
	Adenosine-5′-phosphate	259	15.4×10^3
	Poly-adenosine-5′-phosphate	257	11.3×10^3 [a]
	Deoxyadenosine-5′-phosphate	258	15.3×10^3
	Cytidine-5′–phosphate	271	11.3×10^3
	Deoxycytidine-5′–phosphate	271	9.3×10^3
	Thymidine	267	10.2×10^3
	Guanosine	252	13.7×10^3
Amino acids			
	Phenylalanine (pH = 6)	257	0.2×10^3
	Tyrosine (pH = 6)	275	1.3×10^3
	Tryptophan (pH = 6)	280	5.0×10^3
NAD^+ and $NADP^+$		260	13.4×10^3
	(oxidized)	340	0.00
	(reduced)	340	6.2×10^3
Flavin mononucleotide (flavodoxin)		443	9.1×10^3
		372	7.9×10^3
Cu^{+2} (azurin)		781	3.2×10^3
		625	3.5×10^3
Fe^{+2} – Heme (cytochrome *c*, reduced)		550	27.7×10^3

[a] The decrease in molar absorbance in the polymeric nucleotides compared to monomeric forms is called *hypochromicity*.

spectral lines depends on the vibrational effect of the local environment. The atomic or molecular structures involved in electronic transitions are called *chromophores*. Most of the molecular transitions of biological importance are summarized in Table 28.1. In addition to these organic transitions, important spectra are contributed by various transition metal atoms, such as iron and copper. The common biological chromophores are described in Table 28.1. The transitions of importance in the UV–visible range are those of loosely bound n or π electrons. The absorption maxima of chromophores can undergo changes in frequency maxima, intensity, and width that reflect the structure and interactions of a biomolecule. When a band shifts toward the red or longer wavelengths, the shift is said to exhibit *bathochromicity*. A blue shift is called *hypsochromic*. Changing the length of a region of electron delocalization (polyenes and aromatic hydrocarbons) can be treated like a particle in a box of changing dimension and will exhibit bathochromicity or hypsochromicity if the length of the conjugated system is increased or decreased. These shifts are also seen when the solvent near the chromophore is changed called the *solvent effect*. These effects are summarized in Table 28.2. The intensity of the lines may be increased under certain conditions, and a *hyperchromic* shift is said to occur. Conversely a band that decreases its intensity will experience a *hypochromic* shift. These changes are particularly sensitive to orientation and can be used to study the geometry of macromolecules (Fig. 28.6).

Table 28.2 Causes of red and blue shift of chromophore bands

Hypsochromic shift	—————————————————>				Bathochromic shift
$n \rightarrow \pi^*$ in polar solvent					$n \rightarrow \pi^*$ in non-polar solvent
$\pi \rightarrow \pi^*$ in non-polar solvent					$\pi \rightarrow \pi^*$ in polar solvent
Small-dimension delocalized systems					*Large-dimension delocalized systems*
	Amide bond (190 nm)				
		Nucleic acids (260 nm)			
			β-Carotin (>400 nm)		
				Cis-retinal → trans-retinal	
					Porphyrin (>400 nm)

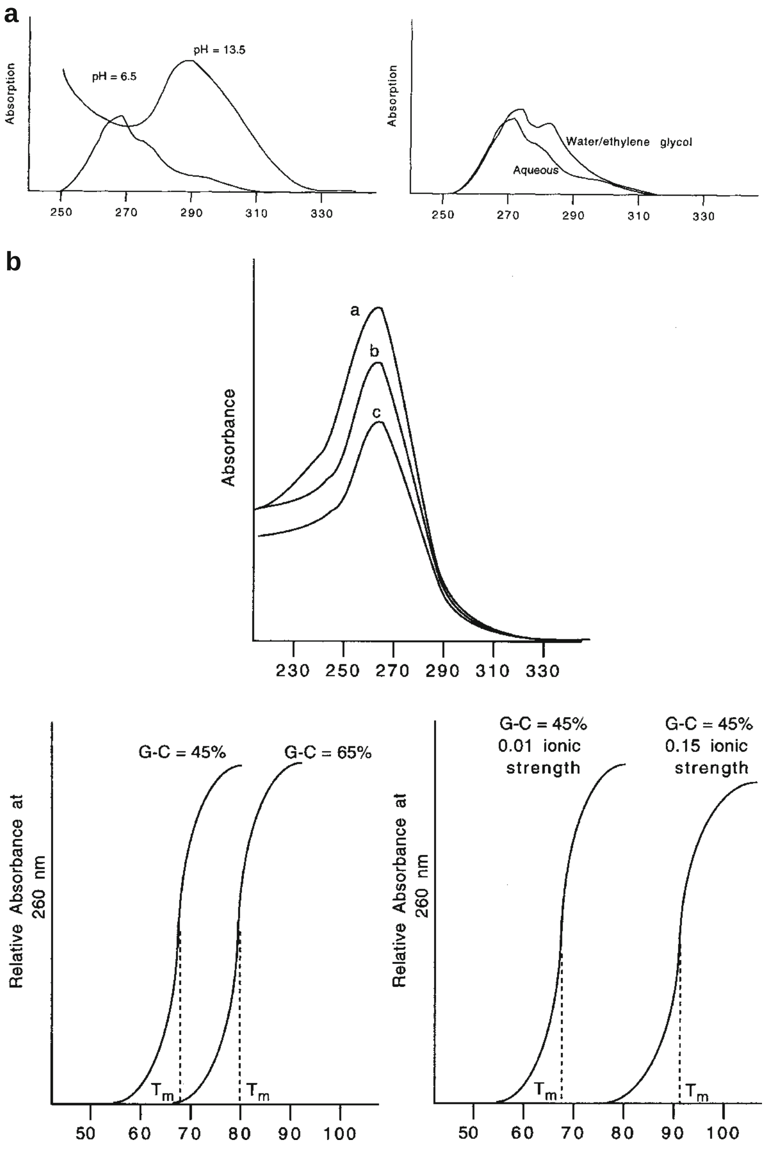


Fig. 28.6 (**a**) The maximum absorbance of tyrosine changes as the side chain undergoes protonation or deprotonation or as the solvent environment is altered. (**b**) Structural studies of biomolecules can take advantage of the sensitivity of the molecular electronic distribution to the environment. Here the absorbance of nucleotide bases and their hypochromicity with increasing structure is deduced in the study of the structure of DNA. In the *top graph* curve *c* is double-stranded DNA, *b* is single-stranded DNA, and *a* is an equal concentration of mononucleotides. The *lower* two graphs show that the T_m (the temperature at which 50% of the DNA is melted or denatured) is dependent on the G–C percentage of bases and on the ionic strength of the solvent. These results are consistent with the stronger stabilization of the DNA structure by the higher association energy of the G–C pair (see Chapter 17). The ionic strength affects the interactions of the strands as mediated by the charged phosphate groups

28.5.1 Absorption Spectroscopy Is a Powerful Tool in the Examination of Dilute Solutions

The great majority of biochemical and biophysical phenomena are performed in solution, and usually spectroscopic studies are performed in dilute solution. Under these conditions the phenomenon of absorption obeys the Beer–Lambert law:

$$I = I_0 e^{-\varepsilon' lc} \quad or \quad I = I_0 10^{-\varepsilon lc} \tag{28.2}$$

where I_0 is the intensity of the incident light, and I is the intensity of the transmitted light after the incident beam passes through a solution of concentration c contained in a 1 cm cell of length, l. The constant ε is called the *extinction coefficient* and has units of L mol-1 cm-1. The intensity of absorption can be expressed as percent transmission ($I/I_0 \times$ 100) or as absorbance:

$$A = \varepsilon lc \tag{28.3}$$

Because only certain transitions are allowed in a given molecule, the extinction coefficient varies with each wavelength of incident light. The measurement of an absorption spectrum is accomplished by guiding a beam of monochromatic light through a sample cell of fixed length (usually 1 cm) containing a solution of the species under investigation (Fig. 28.7). The decrease in intensity of the incident beam is measured electronically by a photomultiplier or a photoresistive circuit. In a scanning spectrophotometer a prism or a diffraction grating is turned stepwise so that a series of incremental bands of wavelengths (determined by the width of a narrow slit, usually giving a range of 2–10 nm) of light are focused onto the sample cuvette in sequential order. The scanning range is determined by a light source that

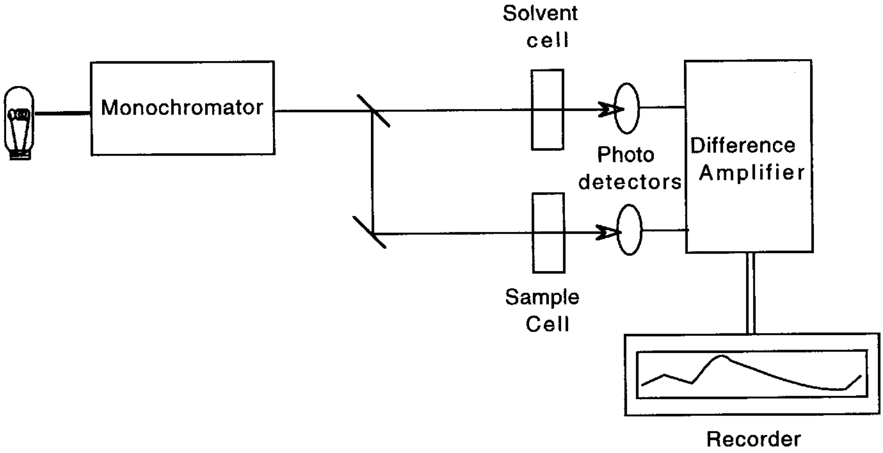


Fig. 28.7 Schematic of a dual beam spectrophotometer

provides a fairly uniform intensity range of wavelengths, usually a tungsten filament for blue through red frequencies (450–900 nm) and a deuterium gas-filled arc lamp for ultraviolet and far blue-violet frequencies (190–450 nm). The absorbance at each wavelength is then recorded with a synchronized recording system. In a dual beam spectrophotometer the incident beam of monochromatic light is, using a beam splitter, split into two beams: one is passed through a reference cell containing only solvent, the other is passed through the sample cell. The parallel beams are compared by circuitry that measures the difference between their intensities and expresses the absorbance or percent transmission directly. A relatively new development in UV–visible spectrophotometers is the use of the optical multichannel analyzer at the detector stage. The incident beam of light is first passed through the reference or sample cell. The diffraction grating or prism then spreads the beam of light onto a CCD (charge-coupled device), which has a series of phototransistors that can read an entire spectra in one unit of time. The array of phototransistors can then be interrogated and the data (photon counts) are read into a computer for analysis. The CCD array is then reset and another cycle of measurement can be taken. This modern spectroscopic tool is extremely valuable in kinetic studies because multiple complete spectra can be recorded in rapid sequence.

28.6 Fluorescence and Phosphorescence Occur When Trapped Photon Energy Is Re-radiated After a Finite Lifetime

When an electron is excited to an orbital with a higher electronic energy, the excess energy is stored. In most cases this energy is transferred to the local environment as heat, and no further radiative exchange occurs. In some molecules the electronic energy states are structured so that the energy can be re-radiated as the electron drops from the excited state back to the ground state. The first case of non-radiative energy transfer is illustrated in Fig. 28.8a. Absorption is dictated by the Franck–Condon principle, and the vibrational energy level into which the electron is excited is within the vibrational energy states available in the ground state. Collisions between solvent molecules and the excited molecule lead to transfer of vibrational energy to the surroundings as heat, and the excited electron falls to lower and lower energy states. Because of the relative configuration between the excited and ground states, the electron can easily make a transition from a vibratory state in the excited system to a nearly degenerate level in the ground state, and it can continue to undergo radiationless energy loss until it reaches the ground state once again. Radiative energy transfer occurs when the electronic configurations between the ground and excited states are arranged as shown in Fig. 28.8b. Again absorption of a photon leads to transition of the electron to an excited state with transfer of some energy to the solvent as heat. However, the lowest vibrational level of the excited state is too far above any vibrational level of the ground state for energy to be relinquished except with radiation of a photon. Re-radiation generally occurs from the lowest vibrational

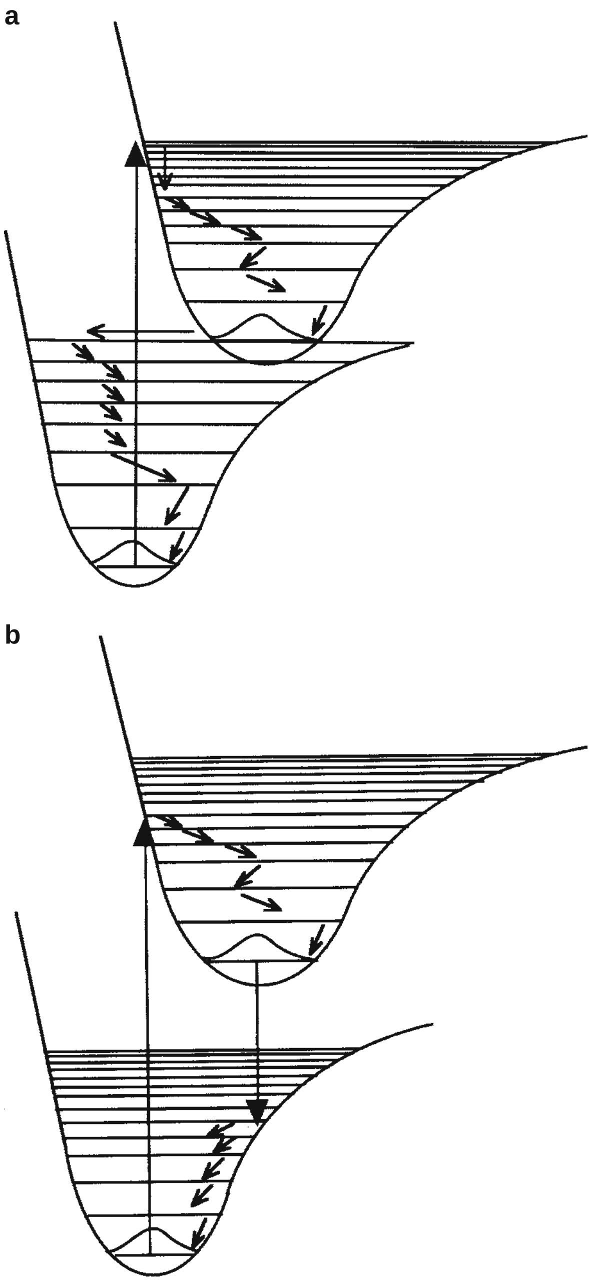


Fig. 28.8 Non-radiative and radiative energy loss. (**a**) The case of non-radiative loss of photoelectronic energy following adsorption. (**b**) In fluorescence, the electron gives up vibratory energy as it falls within the excited energy state. When it reaches the lowest vibratory level in the excited state, this state corresponds to a relatively likely and therefore permitted transition to the ground state. Thus there is the radiation of a photon of longer wavelength than the exciting wavelength

level of the excited state to vibrational levels in the ground state that are consistent with the Franck–Condon rules. The emission wavelength of a fluorescing molecule is independent of the exciting wavelength because the time needed for non-radiative vibratory energy loss is on the order of 10^{-12} s while the lifetime for the excited electronic state is 10^{-9} s. This can be interpreted physically as a situation in which there is plenty of time for every excited molecule to reach the lowest vibratory state before there is any likelihood of re-transition to the ground state. The re-radiation will thus almost always be from the same level to the ground-state configuration. In spite of the long lifetime of the excited state, the time needed for the actual electronic transition is very short (10^{-15} s). The width of the emission spectra therefore depends on the probability of radiative transfer to different vibratory levels in the ground state. The fluorescence spectrum is therefore a characteristic of molecular species and can be used for analytical identification and characterization.

The relatively long lifetime of the excited state and the existence of radiative transfer makes the phenomenon of fluorescence very useful beyond the analytical realm. It is not surprising that many of the excited molecules in a system will not fluoresce. The measure of a molecule's probability of fluorescence following excitation is called the *quantum yield*, ϕ.

$$\phi = \frac{\text{photons fluoresced}}{\text{photons absorbed}} = \frac{\text{rate of fluorescence}}{\text{rate of absorbance}} \tag{28.4}$$

The yield will therefore be dependent on the lifetime of the fluorescent species and whether it gets to re-radiate its energy. Fluorescent studies in which the fluorescent yield is decreased or *quenched* may provide a useful experimental tool to study competing reactions in solution or in biological membranes. A very useful type of quenching is the transfer of the high-energy photon from one chromophore to another. In this process called *excitation transfer*, a donor chromophore absorbs a photon and then fluoresces with the photon transferred to a second chromophore. The fluorescence of the second chromophore depends on its competing reactions and causes for de-excitation; thus a window into its environment can be gained. A very important use of excitation transfer is based on the strong dependence of the transfer efficiency on the intramolecular distance between the two chromophores. The efficiency of this resonance energy or *Förster transfer* depends on the distance as r^{-6} between the chromophores, at least at distances of 1–10 nm. This distance dependence can be used to measure the dimensions within complex macromolecules containing appropriate chromophores (Fig. 28.9). Förster transfer is the basis of the technique in fluorescence microscopy called *FRET* (fluorescence resonance energy transfer) analysis. FRET uses the intensity of a fluorescent signal between a pair of fluorophores as a molecular measuring stick.

While fluorescence is a very practical phenomenon in the daily work of the biological chemist, a related form of photonic de-excitation, phosphorescence, is less practically exploited (unless of course you are a firefly or a lantern fish), but is still of great interest. Phosphorescence is a long-lived excited state that results because the excited state is a triplet state and the ground state is a singlet state. Transitions

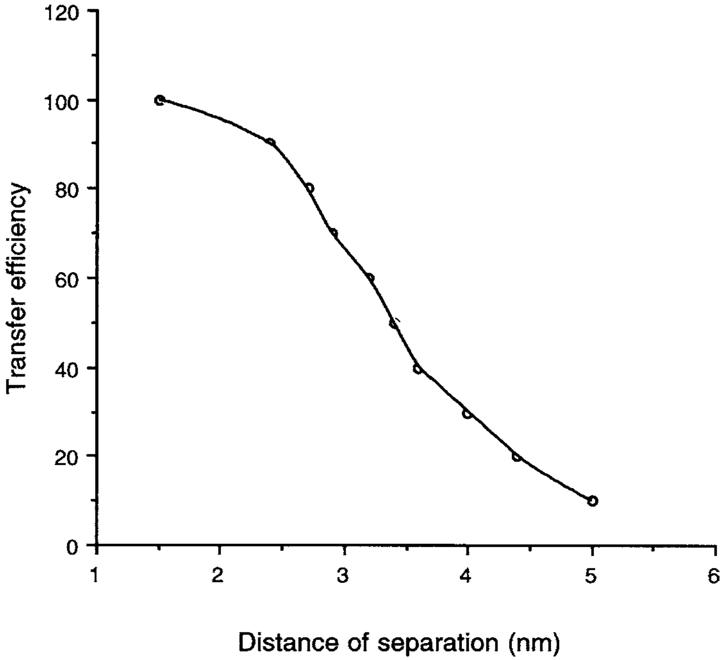


Fig. 28.9 Förster analysis. The linear portion of this curve shows an r^{-6} dependence of the resonance energy transfer on the distance between a pair of naphthyl moieties separated by prolyl residues to form a series of defined geometries [data from Stryer and Haugland (1967)]

between a triplet and singlet state are “forbidden” by the quantum mechanical selection rules and thus the rate of the transition is very low, which accounts for the very long lifetimes. The long lifetimes associated with phosphorescence increase the likelihood of radiationless energy transfer at room temperature, and phosphorescence is more commonly observed at low temperatures. Just as triplet–singlet transitions are forbidden, so are the singlet–triplet transitions necessary to get a phosphorescent system into a triplet state. A system therefore reaches a triplet state because either a chemical reaction occurs that leaves the products in an excited triplet state or an electron that is promoted to an excited state has a degenerate triplet state that is accessible, and it passes over to the triplet state. Phosphorescence can be a competing reaction with an otherwise fluorescent system, thus decreasing the quantum yield of certain systems.

28.7 Electron Paramagnetic Resonance (EPR) and Nuclear Magnetic Resonance (NMR) Depend on Interactions Between Photons and Molecules in a Magnetic Field

In Chapter 27, we reviewed the essential physics of magnetism and examined the generation of the magnetic field. The interaction of this field with moving charges informed our understanding of mass spectrometry. Another important use of magnetism in biochemical studies is the increasing application of electron-paramagnetics (EPR) and nuclear-magnetic resonance (NMR) as a window into structure from the molecular to the level of the organism. The importance of NMR, in particular, has led to rapid development in NMR instrumentation following the first commercial instrument implementation in 1952; machines today are capable of identifying substances present in parts per million concentrations with high resolution and sensitivity. The principles behind NMR have been applied to a variety of investigations, from clinical medicine to polymer and surface science. In the sections that follow, we will examine the production of uniform magnetic fields and the behavior of molecules within such fields.

28.7.1 The Solenoid Shapes the Magnetic Field in a Manner Similar to the Parallel-Plate Capacitor

A solenoid is a coil of wire wrapped uniformly into a long cylinder. Just as the ideal parallel-plate capacitor will set up a constant electric field in space, the solenoid generates a constant magnetic field. Each turn in a solenoid generates a B field. On the inside of the solenoid, the field lines are squeezed tightly together becoming parallel and uniform in space since they cannot cross one another. The net effect is to create a relatively constant magnetic field inside the coil. On the outside of the coil, the magnetic lines are widely spaced, and the magnetic field density is minimal. If the solenoid were infinitely long, the magnetic field outside would be zero. By Ampere’s law, we can find the magnitude of the uniform field inside a solenoid to be

$$\mathbf{B} = \mu_o n I \tag{28.5}$$

where *n* is the number of turns per unit length and *I* is the current through the coil.

28.7.2 Magnetism in Matter Has Distinct Properties

Much of our attention to the interaction of electromagnetic waves with biological matter has focused on the interactions of the electric vector. The interaction of magnetic fields with matter occurs like that of the electric dipole with the electric field: The external electric field orients the dipole parallel to the field with the result that the dipole alignment weakens the field. The movement of electrons in atoms as well as an intrinsic magnetic moment of each electron leads to a net magnetic moment of the atom. When the magnetic dipoles are aligned parallel to an external *B* field, the dipoles will enhance the magnetic field. If the atoms in a material do not have a net permanent magnetic moment, the external field induces a magnetic dipole whose field will oppose the external magnetic field and diminish it. These effects give rise to three types of magnetic interaction between matter and an external magnetic field, *diamagnetism*, *paramagnetism*, and *ferromagnetism*.

Imagine that we have measured the $\mathbf{B}_{\text{ext}}$ of a solenoid and have then placed a core of ice or carbon inside it. If we have a very sensitive measuring device, we can show a change in the $\mathbf{B}_{\text{ext}}$ field on the order of one part in a million. If the core is iron, however, the field is greatly increased (on the order of 10^3). When the core is removed from the solenoid, it behaves as a permanent magnet. Iron is ferromagnetic because it can display magnetic effects when removed from a magnetic field. The other materials only display magnetic properties in the presence of a magnetic field. This response of the bulk material to the magnetic field is called its *magnetization*, **M**. **M** is defined as the net magnetic dipole moment per unit volume and has units of A/m. The net field of the solenoid plus iron core equals the field due to two equivalent solenoids for the iron core, acting as a magnetic dipole, and can be treated as equivalent to a solenoid. The field due to the iron core is then

$$\mathbf{B}_{\text{core}} = \frac{\mu_o NI}{L} = \frac{\mu_o NIA}{LA} = \mu_o \frac{m_{\text{core}}}{V} \tag{28.6}$$

where m_{core} is the magnetic moment of the core and $m_{\text{core}} = NIA$. m_o is the permeability of free space is $\mu_\circ$. The quantity $\frac{m_{\text{core}}}{V}$ is the magnetization and

$$\mathbf{B} = \mathbf{B}_{\text{ext}} + \mu_o \mathbf{M} \tag{28.7}$$

M is written as a vector because it is able to oppose or enhance the external magnetic field, which makes its direction important.

To separate the magnetic field generated from a real external current such as from the solenoid and that generated from the material in the field, the magnetic intensity **H** is introduced. **H** depends on the difference between the net internal field and the magnetization:

$$\mathbf{H} = \frac{\mathbf{B}}{\mu_\circ} - \mathbf{M} \tag{28.8}$$

It follows that Eq. (28.7) can be rewritten as

$$\mathbf{B} = \mu_\circ \mathbf{H} + \mu_\circ \mathbf{M} \tag{28.9}$$

Most materials when placed in a magnetic field develop a magnetization field only when exposed to an external **B** field. The magnetization field either aligns itself parallel or anti-parallel to **H** and varies linearly with respect to **H**. A proportionality constant, χ_{m}, called the *magnetic susceptibility* is the coefficient of this linear relationship:

$$\mathbf{M} = \chi_{\text{m}} \mathbf{H} \tag{28.10}$$

Materials that enhance the field are called *paramagnetic*, and those that diminish it are called *diamagnetic*. Magnetic susceptibilities are positive for paramagnetic materials and negative for diamagnetic materials. The permeability of free space $\mu_\circ$, like the permittivity of free space ε_o, is used when the magnetic field is in a vacuum (i.e., in a space free of magnetic dipoles). We can relate **B** to **H** for any case through the permeability of a material, μ, thus

$$\mathbf{B} = \mu \mathbf{H} \tag{28.11}$$

where μ is defined as

$$\mu = \mu_\circ \left(1 + \chi_\mathrm{m}\right) \tag{28.12}$$

Examination of Table 28.3 shows that most non-ferromagnetic materials essentially appear as a vacuum to a magnetic field, thus indicating minimal interaction of the magnetic field with them.

Table 28.3 Magnetic susceptibilities of a variety of materials

Class	Material	Susceptibility (χ_m at 293 K)
Diamagnetic		
	Water	–9.1 × 10^{-6}
	Carbon (diamond)	–2.2 × 10^{-5}
	Copper	–2.2 × 10^{-5}
	Hydrogen	–9.9 × 10^{-9}
	Carbon dioxide (1 atm)	–2.3 × 10^{-9}
	Nitrogen (1 atm)	–5.6 × 10^{-9}
Paramagnetic		
	Oxygen (1 atm)	2.1 × 10^{-6}
	Magnesium	1.2 × 10^{-5}
	Liquid oxygen (90 K)	3.5 × 10^{-3}
Ferromagnetic		
	Iron (annealed)	5.5 × 10^{3}
	Permalloy (55% Fe, 45 % Ni)	2.5 × 10^{4}
	Mu-metal (77% Ni/16% Fe/5% Cu/2% Cr) 1.0 × 10^{5}	

28.7.3 Atoms Can Have Magnetic Moments

The molecular basis for magnetic effects in matter derives from the appreciation that both electrons and protons are moving charges on an atomic scale and hence represent current. It is the movement of these atomic currents that give rise to the magnetic properties of matter. In terms of bulk magnetic properties, the currents generated by electrons are overwhelmingly more important than the contributions from the protons found in nuclei, and so all the bulk magnetic effects are due to electronic effects. However, it is the proton-derived current that is the basis for nuclear magnetic resonance studies and for magnetic resonance imaging.

Although a quantum mechanical analysis is required for a proper description, all electrons have intrinsic angular momentum. This property is often called the *spin angular momentum* or spin S:

$$S = \frac{h}{4\pi} = 5.2729 \times 10^{-24}\mathrm{Js} \tag{28.13}$$

h is Planck's constant. The intrinsic magnetic dipole moment associated with spin is

$$\mu_{\mathrm{B}} = \frac{e}{m}S = \frac{eh}{4\pi m} = 8.27 \times 10^{-24}\mathrm{JT}^{-1} \tag{28.14}$$

Here e is the charge on an electron and m is its mass. This quantity derived from the spin of the electron is called the *Bohr magneton*, $\mu_{\mathrm{B},}$ and is used as the base unit for atomic and particle magnetic moments.

If we treat the atom as a nearly stationary nucleus around which electrons are moving in a circular orbit, there will be a movement of current around the nucleus. The treatment of this circulating current as a single orbit (Fig. 28.10) occurs because over numerous orbits the net effect is a continuous current ring around the nucleus. Considering a one-electron atom, in which the charge of –e moves in a circular radius r with a period of motion $T = \frac{2\pi r}{v}$, the current will be

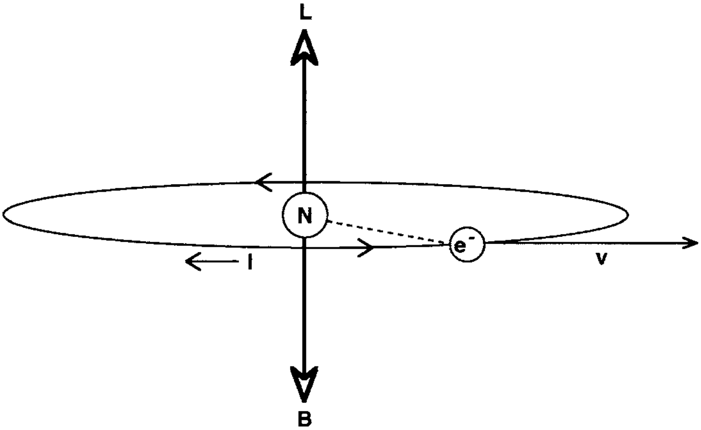


Fig. 28.10 The apparent flow of atomic electronic current leads to a magnetic vector

$$I = \frac{-e}{T} = -\frac{ev}{2\pi r} \tag{28.15}$$

The meaning of the minus sign is that the current is in the direction opposite the movement of the electron. The current loop has an area of πr^2, and thus the orbital magnetic moment can be determined by

$$m_{\text{orbital}} = I\pi r^2 = \frac{ev}{2\pi r}\pi r^2 = \frac{1}{2}evr \tag{28.16}$$

The direction of the magnetic moment is determined as usual by the right-hand rule. For an electron of kinetic energy 1 eV moving in a radius of 10^{-10} m, the m_{orbital} is approximately 5×10^{-24} A m^2. In contrast, a 1 cm^2 macroscopic current loop carrying a current of 10 mA will have a magnetic moment of 10^{-6} A m^2.

The magnetic moment is proportional to the angular momentum carried by the circulating charge; therefore, the magnetic moment is often expressed in terms of angular momentum. We can rewrite Eq. (28.16):

$$m_{\text{orbital}} = \frac{1}{2}evr = \frac{e}{2m_{\text{e}}}m_{\text{e}}vr = \frac{e}{2m}L \tag{28.17}$$

where m_{e} is the electron mass and $L = m_{\text{e}}vr$, the angular momentum of the electron in its orbit. The ratio $\frac{e}{2m_{\text{e}}}$ is a coefficient that links the magnetic moment and the angular momentum and is called the *gyromagnetic ratio*, g_{L}. In the case we have discussed here, $\mathbf{m}_{\text{orbital}}$ and $\mathbf{L}$ point in opposite directions, and so the ratio would carry a minus sign. We can rewrite Eq. (28.17) in vector notation:

$$\mathbf{m}_{\text{orbital}} = g_L\,\mathbf{L} \tag{28.18}$$

The quantum mechanical rules for circular orbits quantize the magnitude of the angular momentum of $\mathbf{L}$ such that $L = \frac{lh}{2\pi}$ where l is an integer. We write the quantized magnetic moment of a single electron atom:

$$m_{\text{orbital}} = \frac{e}{2m_{\text{e}}}l\frac{h}{2\pi} = \frac{eh}{4\pi m_{\text{e}}}l \tag{28.19}$$

Comparison with Eq. (28.14) reveals that $\frac{eh}{4\pi m_{\text{e}}}$ is the Bohr magneton, μ_{B} and so Eq. (28.19) can be rewritten as

$$m_{\text{orbital}} = \mu_{\text{B}}l \tag{28.20}$$

Materials are diamagnetic when each of the electrons in an orbital are paired with one another, one with spin up and one down. In these cases there is zero angular momentum and no permanent dipole moment. Diamagnetism is the effect of an induced moment only and therefore is not temperature dependent. Paramagnetism occurs in atoms with unpaired electrons which therefore possess a permanent magnetic moment. The paramagnetic effect is reduced with increasing temperature, and at a high enough temperature all materials are diamagnetic.

28.7.4 EPR Spectroscopy Allows Exploration of Molecular Structure by Interaction with the Magnetic Moment of an Electron

In earlier sections of this chapter, we restricted our discussion to interactions of the electric portion of the electromagnetic field with the molecule; but the magnetic field can also lead to transitions. The "handle" to which the magnetic field is able to couple and interact is the magnetic field generated by charged particles that possess angular momentum. This description obviously applies to an electron, and electrons do possess magnetic moments. It is correct to posit a pair of magnetic moments for an electron, because a magnetic field is generated both by its intrinsic spin and by its orbital angular momentum. This magnetic field will affect the atomic energy levels and is seen in atomic spectra as fine splitting of spectral lines. Such effects were noted in 1887 by Michelson and Morley, who noted the splitting of the H_a spectral line. If a pair of electrons occupy an orbital, the net spin is zero because one will be $s = +1/2$ and the other will be $s = -1/2$. Under these conditions no net magnetic moment can be detected. However, when the electrons are unpaired, the magnetic moment appears and interaction with the magnetic portion of the electromagnetic field can occur.

The charged atomic particle can orient itself in limited orientations when placed in an external magnetic field (Fig. 28.11). For an unpaired electron placed in a magnetic field on the order of 10^4 Gauss, a resonance condition leading to switching between the allowed orientations can be found in the microwave region of the electromagnetic spectrum. This phenomenon can be used to generate an energy signal and is the basis of *electron paramagnetic resonance spectroscopy* (EPR). In biological systems the principal atoms with unpaired electrons are free radicals, transition metal complexes, and atoms in the triplet state. EPR studies have been used extensively in the study of metalloenzymes such as those containing Cu^{+2} and heme. The transitions between reduced and oxidized form of the metals in these proteins can be detected with great sensitivity and speed with EPR studies. The use of stable free radical species such as nitroxides as "spin labels" for the study of certain physical properties of membranes has been a successful use of EPR spectroscopy. Because the local magnetic environment, derived from the interaction of the unpaired electron with the magnetic moments of certain nearby nuclei, can lead to splitting of the energy levels, a phenomenon called *hyperfine splitting* can be observed in the EPR spectrum. Changes in the hyperfine spectra provide information about the local physical environment of the paramagnetic electron.

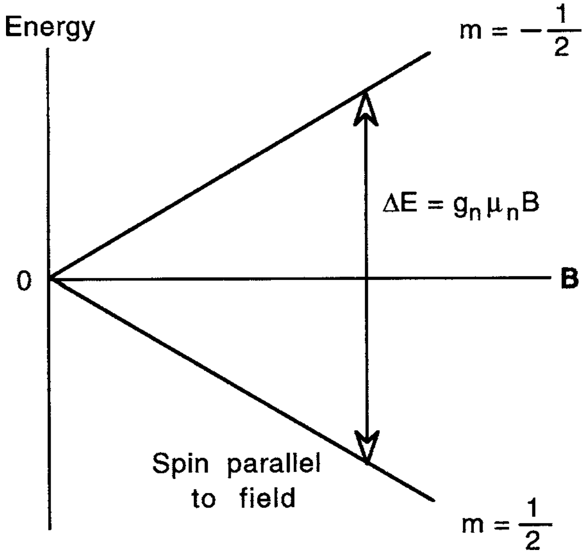


Fig. 28.11 Fundamental physics of NMR and EPR spectroscopy. A "spinning" charge in a magnetic field has two orientations each with different energies. The energy level transition occurs at $\Delta E = g\beta H$. The frequency of adsorption for a photon making this transition is given by $v = \frac{g\beta H}{h}$

28.7.5 *NMR Spectroscopy Employs the Magnetic Properties of Certain Nuclei for Determining Structure*

The electron is not the only atomic component that can generate a magnetic field. A nucleus such as ^{1}H or ^{13}C, which has uneven numbers of protons, also possesses charge and angular momentum. These are the relevant quantum mechanical properties of nuclei required to understand NMR methods. NMR principles are similar to those discussed above for EPR spectroscopy (Fig. 28.11). Magnetic dipole moments are created in rotating charged bodies; the direction of the created dipole is along the axis of rotation. A requirement for NMR detection of a nucleus is that it be both charged and rotate about an axis. The amount a particular nucleus rotates is represented by the nuclear spin quantum number, I. A charged nucleus with a non-zero quantum spin number ($I \neq 0$) will have a magnetic dipole moment. NMR is only sensitive to nuclei with magnetic moments. This includes ^{1}H and ^{13}C each of which have $I = 1/2$ and can be detected with NMR.

The placement of these nuclei in a strong external magnetic field ($\mathbf{B_0}$) causes alignment of those nuclei in limited directions relative to the field. Each direction is associated with a specific energy. The number of orientations possible for a given nucleus is dependent on the spin quantum number:

$$\text{Number of orientations} = 2I + 1 \tag{28.21}$$

^{1}H and ^{13}C can orient themselves in two ways, either with or against the direction of the applied field. It requires less energy for a nucleus to orient its magnetic moment in the direction of the external field; therefore, this orientation has a lower energy than when the magnetic moment opposes the field (Fig. 28.11).

The difference in energy between the two orientations is proportional to external magnetic field strength (B_o):

$$\Delta E = hv = uhB_0 \tag{28.22}$$

where ν is the absorption frequency of a nucleus, h is Planck's constant and μ is the magnitude of the magnetic dipole. This equation gives rise to the Larmor equation (28.23), which via the de Broglie relation provides the direct relationship between the frequency of radiation absorbed (ω) by the nucleus and the external magnetic field.

$$\omega = \frac{\mu B_0}{I} = \gamma B_0 \tag{28.23}$$

γ is called the gyromagnetic ratio and is the ratio of the magnetic moment to the angular momentum of a particular nucleus (μ/I). γ for the ^{1}H nucleus is 42.58 MHz/T and for the ^{13}C nucleus is 10.705 MHz/T. Absorption of radiation causes a nucleus to a flip its magnetic moment, meaning the moment moves from being aligned with the magnetic field to being aligned against the field. A nucleus is considered to be in its excited state when its magnetic moment is aligned against the magnetic field, thus absorption causes a nucleus to move from the ground state to the excited state. A nucleus can excite and relax cyclically through absorption and energy loss. This is called *precession*. A nucleus alternating or precessing between the excited and the ground states is said to be in a state of resonance.

NMR spectroscopy is able to utilize these nuclear properties to derive information about the identity of an organic molecule. This is familiar as the Boltzmann relationship with the energy difference for Eq. (28.22) in the exponent. In NMR, samples are placed in a uniform magnetic field, thus causing the nuclei with magnetic moments to align their spins with or against the magnetic field as described above. The sample is then exposed to electromagnetic radiation in the range of radio wave frequencies (1–300 MHz). The frequency at which the nuclei absorb energy is detected, thus providing information about the chemical environment and connectivity of the sample molecule (Fig. 28.12).

The ratio of nuclei that align with the magnetic field (+1/2 spin) in the ground state to nuclei that align against the field (–1/2) in the excited state is given by the following relationship:

$$\frac{N^*}{N} = e^{-\mu h B_0/kT} \tag{28.24}$$

where $N^*$is the number of nuclei in the excited state, N is the number of nuclei in the ground state. In NMR experiments this ratio is always close to 1, meaning the ratio of nuclei in the excited state is nearly equal to the number in the ground state. It is only possible to detect nuclear transitions that begin in the ground state; therefore, the sensitivity of NMR is very low. When this ratio is equal to one,

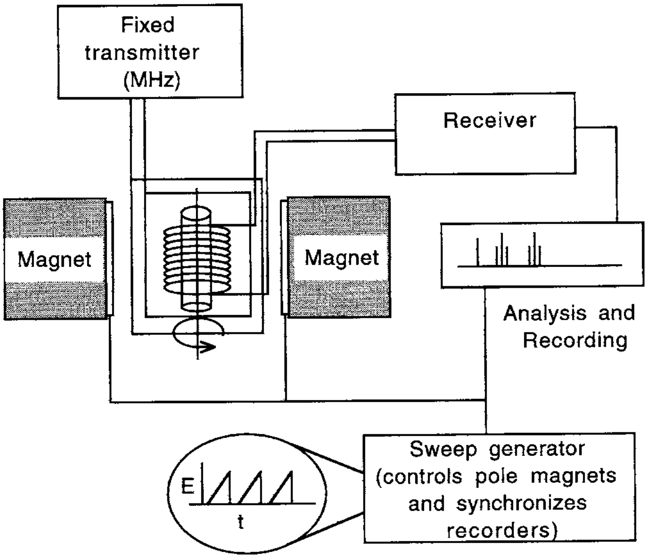


Fig. 28.12 Schematic of the NMR spectroscope: In magnetic resonance studies the material under investigation is placed in a high-field strength magnet (*H*). Either a fixed field magnet of ≈ 10^4 Gauss is used with a scanning of the microwave spectrum (MHz) or a fixed microwave frequency is applied with the magnetic field scanned. A power drop at the transmitter is measured with absorption of the microwaves or a power increase is measured with emission of the photons at the resonance point of the system

there are no excess molecules in the ground state to excite, and no NMR signal is detected. Increasing the applied magnetic field strength and decreasing the temperature will maximize this ratio, thus increasing signal intensity. Current efforts to increase NMR sensitivity focus on developing high field strength magnets.

As shown above, the frequency at which a given nuclei resonates (ω) depends on the magnetic field to which it is exposed. In theory, each nucleus should be exposed to the full strength of the applied magnetic field; however, in reality, the effective magnetic field a nucleus experiences depends on the environment in which that nucleus exists. For NMR, this means that nuclei (e.g., ^{1}H or ^{13}C) in different chemical environments will absorb at slightly different frequencies due to the different adjacent atoms.

The differences in effective magnetic field strengths are due specifically to the variety in electron density for given atoms. When electrons are placed in a magnetic field, they will rotate and generate a small magnetic field that opposes the larger applied magnetic field. This opposing field acts as a shield for the nucleus from the applied field and will decrease the effective magnetic field as follows:

$$B_{\text{eff}} = B_0 - \sigma B_0 \tag{28.25}$$

where σB_0 is the net shielding by the drift of local electrons and σ is the diamagnetic shielding constant. Changes in chemical environment will lead to changes in the electron density surrounding a given nucleus, which will change the σB_0 and thus the frequency at which a given nucleus will absorb. The degree of shielding is given a parameter σ which is reflected in the resonance condition:

$$\text{without shielding}: \quad v = \frac{2\mu_{\mathrm{m}} H}{h} \tag{28.26}$$

$$\text{without shielding}: \quad v = \frac{2\mu_{\mathrm{m}} H\,(1-\sigma)}{h} \tag{28.27}$$

where μ_m is the nuclear magnetic component in the direction of the field H.

The degree of shielding leads to magnetically distinct nuclei that can be identified by measuring the shift of the resonance condition either toward a higher magnetic field (upfield) or toward a lower field (downfield). The greater the shielding, the greater the magnetic field needed to achieve the resonance state. These movements are called *chemical shifts* (δ) and are measured in parts per million (ppm) versus a reference or standard molecule that are added in each experiment. The dominant effect reflected in the chemical shift is the electron density at the nucleus. Downfield shifts (increasing δ) indicate a decreasing electron density such as would occur when the nucleus is adjacent to a strongly electronegative (electron withdrawing) group. For example, a methyl proton next to another methyl group has a δ of ≈1 ppm while the acidic proton in a carboxylic acid has a δ of ≈ 10 ppm. Table 28.4 lists the chemical shifts found in a variety of different protons. The chemical shift can be used to sense local changes in conformation in a macromolecule. Movements of

Table 28.4 Chemical shifts for protons in bio-organic molecules

	Proton molecular environment					
δ(in ppm)	Methyl	Methylene	Methine	Olefinic	Aldehydic	Aromatic/ Others
0	$(CH_3)_4Si$					
0.22		Cyclopropane				
0.92	$C(CH_3)_4$					
1.17	CH_3CH_2OH					
1.44		Cyclohexane				
2.07	CH_3COCH_3					
3.38	CH_3OH					
3.59		CH_3CH_2OH				
3.95			$(CH_3)_2CHOH$			
4.6				$(CH_3)_2C{=}CH_2$		
5.57				Cyclohexene		
7.27						Benzene
7.73						Naphthalene
9.72					CH_3CHO	
9.96					C_6H_5CHO	
12						Imino groups

electronegative groups, aromatic rings, or paramagnetic species in the vicinity of a proton under study can have dramatic effects on the δ, thus providing a window into the conformational changes. Electron shielding is also relevant to ^{13}C atoms; therefore, ^{13}C NMR spectra also provides information about carbon functional groups.

In ^{1}H NMR, the peaks associated with each functional group are proportional to the number of protons located in that chemical environment. Therefore, the peak areas provide information about the relative number of hydrogens located within that environment. These relative numbers are the *integration constants* of each peak. Examination of the relative integration constants in a particular spectrum is a useful tool for determining the precise connectivity of a molecule. This approach is applicable to ^{13}C NMR as well though the integration constants are reflective of ^{13}C atoms.

In proton NMR, the effective field strength depends not only on the electron density of neighboring atoms but also the neighboring protons. These hydrogen nuclei will spin in the magnetic field either with or against the field direction and will either serve to increase or decrease B_{eff}. This change in the effective field affects the magnetic environment of the neighboring carbon's hydrogens altering the frequency at which the affected nuclei will absorb. This change in effective field is called *spin–spin splitting*. Spin–spin splitting causes the peaks for a given proton to move from the original frequency (due to chemical environment) to a slightly higher and lower frequency. The frequency shift that spin–spin splitting moves the reference proton peaks is a constant for a given molecular conformation. This frequency shift is referred to as the *J-coupling* constant.

Spin–spin splitting is generally strong only between protons on adjacent carbons. The number of fine peaks associated with a given proton, therefore, depends on the number of adjacent hydrogen atoms. When the neighboring hydrogen atoms have the same chemical environment as the proton of interest and the separation between the chemical shifts of sets of magnetically equivalent protons is sufficient (i.e., if va – vb>10 J), then each set of n equivalent protons will split their neighboring protons into a multiplet of (n +1) lines. When the neighboring hydrogens are in an environment different from the reference proton, the number of peaks seen can be calculated as follows: $(n_a + 1)\ (n_b + 1)$, where n_a and n_b represent the number of type A neighboring hydrogen atoms and number of type B neighboring hydrogen atoms, respectively. The multiplet pattern caused by hydrogen coupling can provide useful information about the connectivity of a molecule, as the pattern of a multiplet is a direct result of the environment in which the reference proton resides.

The relative size or intensity of the multiplet lines corresponds to the coefficients of a binomial expansion. The patterns expected are summarized in Table 28.5. As the separation between the chemical shifts of the sets of protons decreases, the analysis becomes more complicated and the multiplets approach one another with increasing intensity of the inner lines and loss of signal in the outer lines. Spin–spin coupling is independent of the applied magnetic field while the chemical shift is not. Therefore, a higher frequency spectrometer will increase the chemical shift

Table 28.5 Spin–spin coupling patterns

Number of protons	Multiplet type												
0							1						
1						1		1					
2					1		2		1				
3				1		3		3		1			
4			1		4		6		4		1		
5		1		5		10		10		5		1	
6	1		6		15		20		15		6		1

separations to a degree sufficient to maintain the simple first-order analysis outlined above. Improved instrumentation thus can result in increased resolution of the structural information available from the NMR study. For conformational studies, *J* can be very sensitive to the relative position of protons on adjacent carbon atoms. *J* is proportional to the dihedral angle of a CH_3–CH_3 grouping varying from 0 to 180° (*cis* orientation = 0° and *trans* orientation = 180°). The dependence is given by the Karplus equation:

$$J = \mathrm{A} + \mathrm{B}\cos f + \mathrm{C}\cos 2f \tag{28.28}$$

The coefficients must be either calculated or taken from empirical measurements but it can be seen that *J* will be maximal at 0 and 180° (usually around 10 Hz) and almost 0 Hz at 90°. It should be appreciated that substantial information can be deduced from NMR spectra relating the identity as well as configuration and conformation of the physical state space of protons as well as other NMR active molecules.

28.7.6 Further Structural Information Can Be Found by NMR Studies of Nuclei Other Than Protons

We have explored the ways in which different structural information can be derived from proton NMR. Although proton NMR is a very useful tool when trying to predict molecular structure, it is even more powerful when paired with other types of NMR investigations. The way in which these different types of spectra are collected has to do with different pulse sequences or the pattern of radio frequency pulses that are applied to a sample to induce precession. We will review several NMR techniques to provide a flavor of this complex technique.

^{13}C NMR provides several advantages that proton NMR does not. First of all, ^{13}C NMR has a much larger range of chemical shifts in comparison to proton NMR (200 versus 10 ppm), which decreases the overlap of carbons in different environments and makes it easier to determine the number of different carbon environments within a molecule. These spectra are also easier to interpret, as adjacent ^{12}C does not induce spin–spin splitting and it is highly unlikely that two ^{13}C will be next to each other in a molecule. ^{13}C NMR is often the best way to first approach structure determination because of the simplicity of the spectra; analysis of these spectra provides the experimenter with an approximation of the different types of functional groups present, which narrows the possible interpretations of proton spectra.

A more specialized type of ^{13}C NMR is the distortionless enhancement of polarization transfer (DEPT). DEPT spectra provide information about the hydrogen environment of a particular carbon. More specifically, the different carbon environments are still represented at their respective chemical shifts: CH and CH_3 typically appear as normal peaks, CH_2 peaks appear inverted, and carbons in a quarternary environment are not visualized. These spectra are obtained through a series of radio wave pulses with various time delays, the details of which can be found in the references at the end of this chapter. DEPT spectra make it easier to determine the potential carbon environments of a molecule by eliminating functional group possibilities at each chemical shift.

Although these NMR techniques can provide a wide variety of molecular information, they are one-dimensional spectra that can become very complicated when dealing with complex organic and biological molecules. The development of two-dimensional NMR has allowed experimenters to gain more information from one-dimensional spectra by simplifying the spin–spin splitting interpretation.

One two-dimensional technique that is useful is the homonuclear COSY (COrrelational SpectroscopY) experiment. These experiments use particular pulse sequences to map the proton–proton coupling in a sample molecule. These pulse sequences cause an interaction of magnetization between coupled hydrogens. These interactions are then plotted on a two-dimensional COSY spectrum with the one-dimensional spectrum on both the *x*- and *y*-axes. The presence of an intersection indicates spin coupling between two hydrogen nuclei and, thus, provides unequivocal information about molecule connectivity. A second related technique is heteronuclear multiple quantum correlation (HMQC). This technique provides information about carbon to hydrogen connectivity; in these spectra, the ^{13}C spectrum is on one axis and the ^{1}H spectrum is on the other.

Recently two other NMR techniques, relaxation and the nuclear Overhauser effect have become very important. While the NMR signals that we have been describing are generally obtained by measuring the absorption of energy by the molecules under study from the incident radio waves, information can also be obtained by intensely irradiating the sample thus saturating it. We consider first the relaxation studies. If enough radio energy is applied to the sample, both the lower and the upper energy levels are equally occupied. The signal is thus saturated and an equal number of emissions and absorptions leads to a steady state condition. Once the upper energy level is saturated there will be a spontaneous emission of energy to return to the equilibrium population levels of the energy levels for the nuclei in the magnetic field. The kinetics of this return to the equilibrium are first order with a relaxation time given by a time constant called the *spin–lattice relaxation time*, T_1. The measurement of T_1 can be accomplished by first saturating a sample with radio energy (the NMR peak area will be 0) and then removing the radiation and listening with a radio receiver to the emission photons. The peak area measured against time will follow the exponential kinetics of a first-order process and yield T_1:

$$M = M_{\mathrm{eq}}\left(1 - e^{t/T_1}\right) \tag{28.29}$$

M is the magnetization (what we are measuring with the peak area) while M_{eq} is the magnetization measured at equilibrium. The significance of the T_1 relaxation time is that it depends on the local environment (the lattice) and thus is sensitive to both local magnetic effects, like a paramagnetic species and as well to the exchange of chemically similar protons between the environment and the irradiated sample. Information about the kinetics of chemical exchange and the influence of the environment are therefore available from T_1 time measurements. A second relaxation

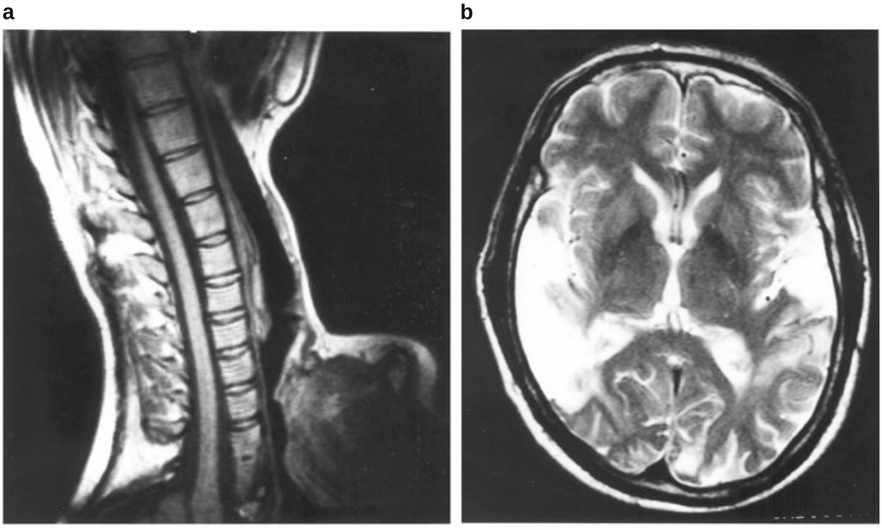


Fig. 28.13 Magnetic resonance images of the brain and spinal cord. In image (**a**), a mid-sagittal section of the neck is shown in a T_1-weighted image. The MRI is tuned to the resonance echo of water protons. On T_1 images, bulk water-like cerebrospinal fluid (CSF) is black because it has no significant signal, normal cellular tissue is gray appearing with a moderate signal and fat has a bright (*white*) signal. From *left* to *right*, tissue planes are seen: skin and subcutaneous fat, muscle and spinous processes of the spinal column, *black* CSF, *gray* spinal cord, *black* CSF, the vertebral bodies of the spinal column with the lens-shaped intervertebral discs, the air-filled trachea, and the front of the neck. Image (**b**) is a T_2-weighted axial section through the middle of the brain. The CSF is *bright white* in T_2 images in contrast to the absence of signal for this material in T_1. The outlines of CSF surrounding the brain can be seen around the edges of the study but the large bright region on the *left* is a region of brain infarction due to a stroke in this patient. The stroke causes swelling or edema which is actually a collection of body water in the tissues. This property of T_2 imaging makes this an extremely sensitive technique for identifying damaged tissues

time constant called T_2 is shorter than T_1 and represents the spin–spin relaxation time. This time constant is independent of the environment and occurs because of interactions between the spins of equivalent nuclei.

The relaxation effects are useful for the study of kinetic processes but their most dramatic application has been in the field of medicine. Combining the saturation methods with computer reconstruction methods, the T_1 and T_2 times of water protons can be utilized to provide amazingly detailed pictures of the soft tissues and associated disease causing processes. T_1-weighted images of the brain and spinal cord look like the actual tissues themselves and a T_2-weighted image is highly sensitive to the bulk water seen normally in the cerebrospinal fluid and also in pathological processes such as the edema or tissue swelling associated with strokes and tumors (Fig. 28.13).

Finally, the saturation technique can be used to saturate an NMR peak of one type of proton and observe how this saturated population affects a nearby peak. The interaction of a saturated population with another NMR peak leading to a change in intensity is called the nuclear Overhauser effect and can be used to localize close neighbors. The effect is dependent as r^{-6} and therefore structural determinations can be made on the order of less than 0.5 nm. Thus structural determinations can be made on the order of that associated with x-ray crystallography but on macromolecules in solution. The coupling of NOE spectroscopy with Fourier transform techniques has lead in recent years to significant advances in macromolecule structural determination. Such structural determinations are now being deposited in protein databanks along with x-ray crystal structures.

Further Reading

History

Thomas N.C. (1991) The early history of spectroscopy, *J. Chem. Educ.*, **74**:65–67.

General

Freifelder D.M. (1982) *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*, 2nd edition, W.H. Freeman, New York.

Kettle S.F.A. and Norrby L.J. (1990) The Brillouin zone – an interface between spectroscopy and crystallography, *J. Chem. Educ.*, **67**:1022–1027. (This explores an intersection of a subject matter from the different perspectives of a crystallographer and a spectroscopist. It is a nice demonstration of how the same natural system maps differently into two abstract spaces).

Lykos P. (1992) The Beer–Lambert law revisited, *J. Chem. Educ.*, **69**:730–732.

Macomber R.S. (1997) A unifying approach to absorption spectroscopy at the undergraduate level, *J. Chem. Educ.*, **74**:159–182.

Nölting B. (2006) *Methods in Modern Biophysics*, 2nd edition, Springer-Verlag, Berlin.

Serdyuk I.N., Zaccai N.R., and Zaccai J. (2007) *Methods in Molecular Biophysics*, Cambridge University Press, Cambridge.

Sheehan D. (2009) *Physical Biochemistry: Principles and Applications*, 2nd edition, Wiley-Blackwell, Oxford.

Stryer L. (1987) The molecules of visual excitation, *Sci. Am.*, **257, 1**:42–50.

Van Holde K.E., Johnson W.C., Ho P.S. (2005) *Principles of Physical Biochemistry*, 2nd edition, Prentice Hall, Upper Saddle River, NJ.

Fluorescent Spectroscopy

Stryer L. (1978) Fluorescence energy transfer as a spectroscopic ruler, *Ann. Rev. Biochem.*, **47**: 819–846.

Stryer L. and Haugland R.P. (1967) Molecular rulers, *Proc. Natl. Acad. Sci. U S A*, **58**:719–726. (The original application of Förster energy transfer to the purpose of intermolecular measuring.)

NMR and MRI in Biological Applications

Beaulieu C. (2002) The basis of anisotropic water diffusion in the nervous system – a technical review. *NMR Biomed.*, **15**:435–455.

Bryant R.G. (1996) The dynamics of water-protein interactions, *Annu. Rev. Biophys. Biomol. Struct.*, **25**: 29–53.

Dyson H.J. and Wright P.E. (1996) Insights into protein folding from NMR, *Annu. Rev. Phys. Chem.*, **47**:369–395.

Ernst R.R., Bodenhausen B., and Wokaun, A. (1992) *Principles of Nuclear Magnetic Resonances in One or Two Dimensions*, Oxford University Press, Oxford.

Jellison B.J., Field A.S., Medow J., Lazar M., Salamat M.S., and Alexander A.L. (2004) Diffusion tensor imaging of cerebral white matter: A pictorial review of physics, fiber tract anatomy, and tumor imaging patterns. *Am J Neuroradiol.*, **25**:356–369.

Raichle M.E. (1994) Visualizing the mind, *Sci. Am.*, **270, 4**:58–64.

Sanders J.K.M. and Hunter B.K. (1993) *Modern NMR Spectroscopy: A Guide for Chemists*, 2nd edition, Oxford University Press, Oxford. (Covers the principles for understanding conformational studies with NMR.)

Problem Sets

1. You are studying an unknown compound and find that it is soluble in water and in methanol. You perform UV–visible absorption spectroscopy. In water there is a small adsorption peak at 275 nm that shifts to 290 nm when the same scan is performed in methanol. What is the likely structure and the electronic transition that gives rise to this observation?
2. Calculate the concentration of NADH ($e = 6200$). The absorbance is measured at 340 nm in a 1 cm cell to be (a) 0.345; (b) 0.556; (c) 0.145.
3. You are studying an ion channel for which you want to know the opening pore dimension. Fortunately there are two and only two sites at the mouth of the pore

to each of which a fluorescent label may be attached. You decide to study the dimension with Forster analysis, the efficiency of Forster transfer is given by

$$\text{Efficiency} = \frac{1}{1 + (r/R_0)^6}$$

You choose a pair of fluorochromes whose characteristic R_0 value in water is 27 Å. You measure an energy transfer efficiency of 92%. What is the approximate dimension in the pore.

4. A solvent with an ionic strength of 0.15 alters the R_0 of fluorochrome pair above to 15 Å because of the effect of the local electric field from the ions on the resonance transfer. You reconstitute your pore assembly and notice that when the conductivity of the membrane increases the Forster efficiency drops to 25%. Explain this effect and describe the state of the environment in the pore in the conducting and non-conducting state.
5. The ratio of NAD_+ and NADH can be used to spectrophotometrically measure the redox potential of a system that includes this pair. What is the redox potential of the following samples:

   (a) absorption at 260 = 0.345; absorption at 340 = 0.345.
   (b) absorption at 260 = 0.105; absorption at 340 = 0.345.
   (c) absorption at 260 = 0.965; absorption at 340 = 0.250.

6. You are studying the effect of superoxide anion production by macrophages (one of the white cells in the immune system) on membrane structure. A nitroxide EPR spin label is added to the membrane. EPR monitoring of the spin-label signal is continuous. As the experiment starts the first change noted is the development of hyperfine splitting of the spin-label signal. Account for this change.
7. The experiment from question 6 continues. You continue to study the effect of superoxide anion production by macrophages on membrane structure. A nitroxide EPR spin label is added to the membrane. EPR monitoring of the spin-label signal is continuous.

   (a) The signal is noted to broaden. How can you interpret this change?
   (b) What other methods could be used to prove your explanation for (a)?

Chapter 29
Molecular Structure from Scattering Phenomena

Contents

29.1 The Interference Patterns Generated by the Interaction of Waves with Point Sources Is a Valuable Tool in the Analysis of Structure

We noted in Chapter 7 that when waves interact there will be interference that is either constructive or destructive. All waves, whether of sound, water, or light, will exhibit spatial interference patterns. If two point sources are each generating coherent (in-phase) waves, patterns of brightness and darkness will be seen coinciding with the constructive and destructive interference of the waves, respectively (Fig. 29.1). The reason for the interference pattern can be physically appreciated by an experiment in which a single source of sine waves (such as a sound) is played into two tubes of different length that terminate at a microphone that displays the arriving waves on an oscilloscope. At the beginning of the tubes, each wave is in phase with the other. But though the waves start in phase, when they reach the ends the tubes, because of the different path lengths, they will not necessarily still be in phase. If the two paths are equal, then the wave peaks will reach the microphone at

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_29, © Springer Science+Business Media, LLC 2010

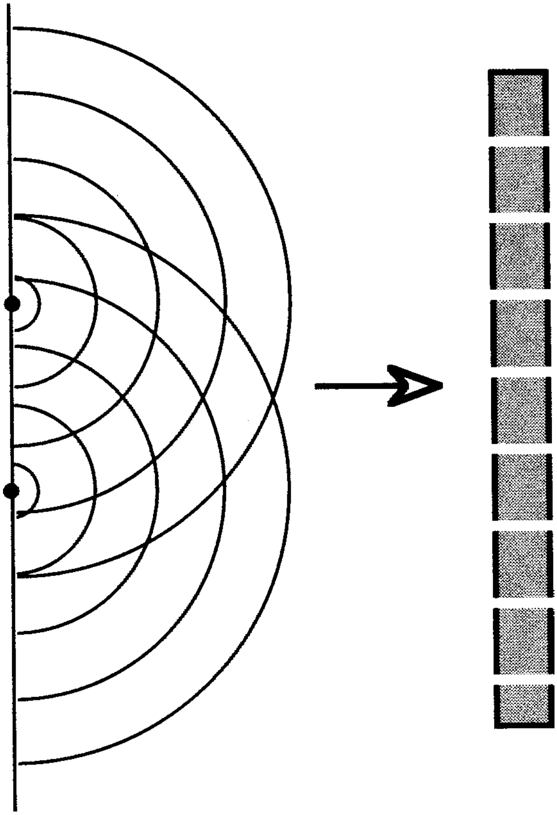


Fig. 29.1 The pattern of constructive and destructive interference generated by two point sources

the same time, and they will interfere constructively. If the length is different, then interference will also occur, but whether it is destructive or constructive will depend on the difference in length $L_2 - L_1 = \Delta L$ of the path. If $\Delta L = \lambda/2$ then a wave peak and trough will reach the microphone simultaneously, the phase difference will be 90° and the interference will be completely destructive. Any length in between will give an intermediate amplitude. For the two extreme conditions of interference:

$$\text{Constructive} \qquad \Delta L = n\lambda, \qquad (n = 0, \pm 1, \pm 2, +3, ...) \tag{29.1}$$

$$\text{Destructive} \qquad \Delta L = \left(n + \frac{1}{2}\right)\lambda, \qquad (n = 0, \pm 1, \pm 2, +3, ...) \tag{29.2}$$

It should be clear that this experiment gives the same result as would be expected from two point sources emitting a coherent harmonic wave with observers placed at two different distances from the sources.

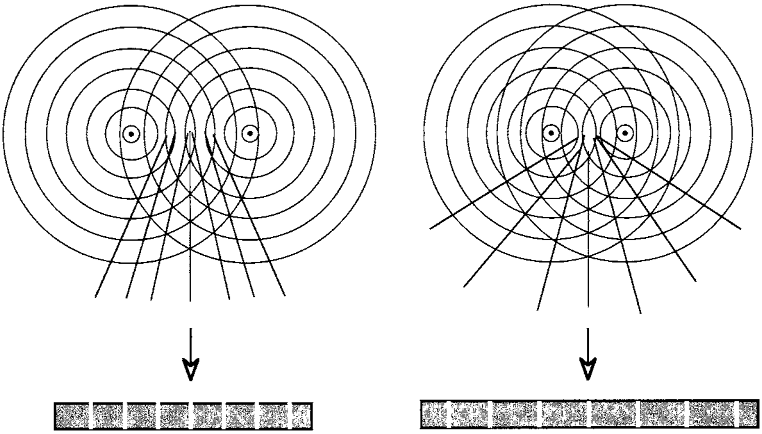


Fig. 29.2 Pattern of maxima and minima for differing distances between the point sources. The lines passing through the nodes are called the orders of diffraction. The *arrow* is in the line of the zero-order diffraction. First-, second-, and third-order diffraction can also be seen in this figure

Look at Fig. 29.2, which shows the maxima and minima for two point sources separated by different distances. If we know where the lines of the maxima and minima occur, we can calculate the distance separating the two point sources. The positions of these lines are straightforward:

$$\text{for maxima} \qquad d \sin\theta = n\lambda \qquad (n = 0, \pm 1, \pm 1; 2, +3, ...) \tag{29.3}$$

$$\text{for minima} \qquad d \sin\theta = \left(n + \frac{1}{2}\right)\lambda \qquad (n = 0, \pm 1, \pm 2, +3, ...) \tag{29.4}$$

Earlier we saw that when we look at propagating wavefronts from a great distance away (with respect to the wavelength), the waves can be treated as if they are plane waves. A screen placed at this distance will show a pattern of interference with alternating bright and dark lines representing the positions of the maxima and minima, respectively. Measuring the spacing between the lines and knowing the distance from the point generators to the screen we can use Eqs. (29.3) and (29.4) to calculate the spacing between the generators. (This analysis will allow us to explore the dimensions of regular structure.) Figure 29.3 shows the geometry needed for these measurements. The intensity of the pattern can be derived from the equations for the superposition of waves, [Eqs. (7.112) and (7.113)], and the geometry as described. The amplitude of the interference pattern with respect to x is

$$A = 2A_0 \cos \frac{1}{2} \frac{\pi x d}{\lambda D} \tag{29.5}$$

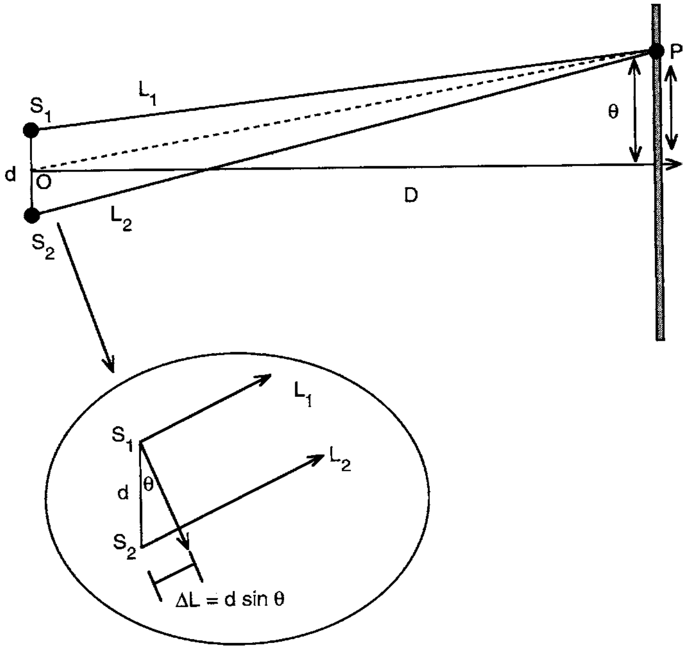


Fig. 29.3 The patterns of the maxima and minima of two point sources can be calculated from a geometric treatment as long as the wavelength, λ, and the separation of the sources, *d* is given. The geometric construction is shown. For convenience an origin, O, is defined as the midpoint between the two point sources, S_1 and S_2. We add a centerline (*D*), which is perpendicular to and bisects *d*. We measure a distance *R* from O to point P, which is on our "interference" screen. The angle θ is the angle that OP makes with respect to the centerline. We are interested in two lengths, L_1 and L_2 from each point source to *P*. The difference between L_1 and L_2 is ΔL and $\Delta L = d \sin\theta$. When OP falls on the centerline, ÆL is 0: the centerline is everywhere an interference maximum. The phase difference at point *P* is simply $\frac{2\pi}{\lambda}\Delta L$. The distance, *x*, measured along the screen from the central point (at the intersection of the centerline and the screen) and *P* is related to $x = D\tan\theta$ by the relation $x = D\tan\theta$. Because θ is very small in cases of interest, $\sin\theta = \tan\theta = \theta$. Thus, we can write $\sin\theta \approx \frac{x}{D}$. We can measure the distance between maxima, and we know the distance from the point to the screen, so we can calculate $\sin\theta$. We then use the equations in the text to find *d*, the spacing between the sources

and the intensity is the square of this

$$I = 4A_0^2 \cos^2 \frac{1}{2} \frac{\pi x d}{\lambda D} \tag{29.6}$$

A plot of the intensity versus *x* is shown in Fig. 29.4. A similar analysis can be done for three, four, or more coherent point sources. This is a good exercise to do until you are convinced that as more and more coherent sources are added, the secondary maxima become weaker and weaker and the major maxima become stronger. The spacing and intensity of these patterns forms the basis for x-ray diffraction and related methodologies.

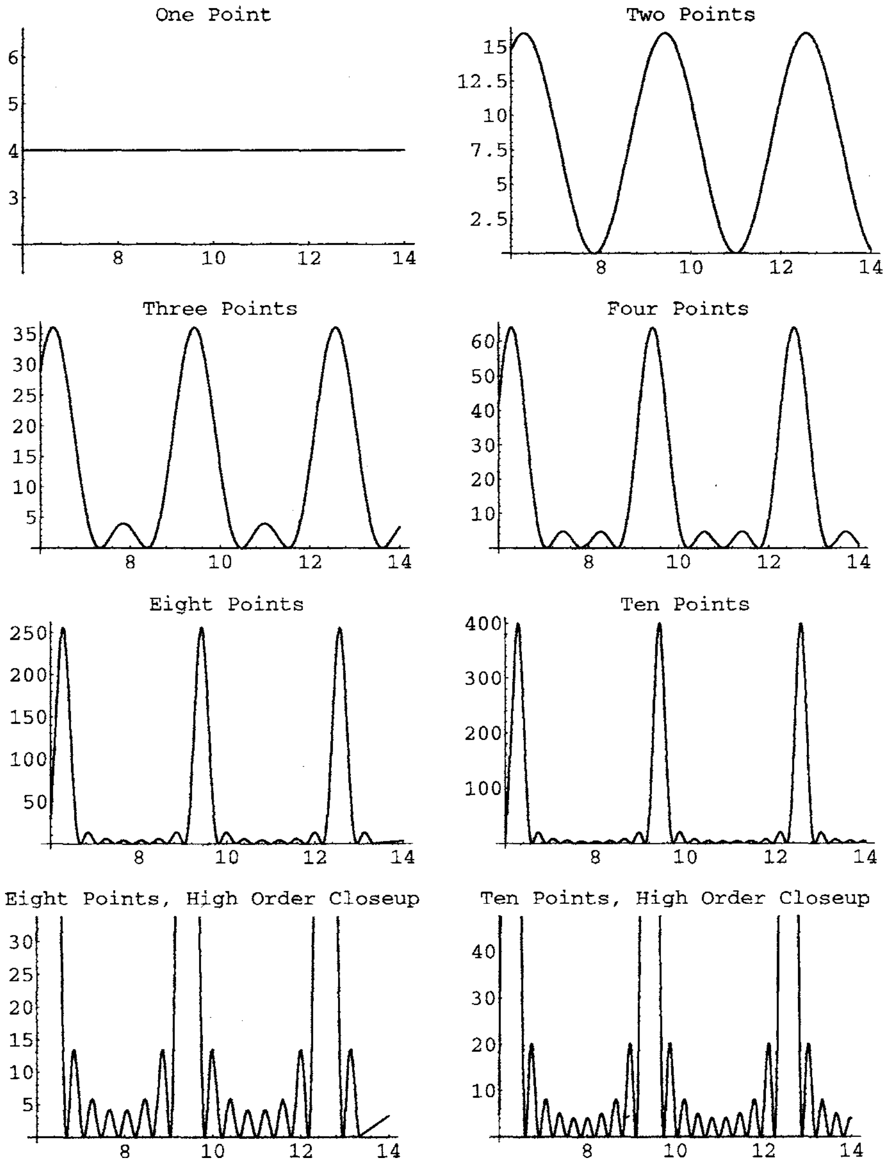


Fig. 29.4 A plot of the intensity pattern for a variety of number of coherent interfering sources

29.2 Diffraction Is the Result of the Repropagation of a Wave

It is easy to describe the propagation of an unobstructed wave. When a wave strikes an obstruction the propagation is considerably more complex. Figure 29.5 shows the wave that propagates beyond an obstruction with a very small opening. The small

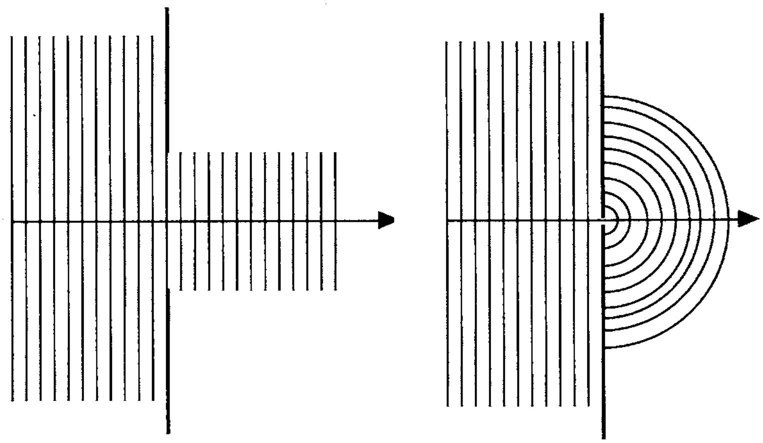


Fig. 29.5 Comparison of wave propagation through obstructions of different sizes. *Left:* The propagation of a wave through an opening larger than the wavelength of the incident wave. *Right:* The propagation of a wave through an opening smaller than the wavelength of the incident wave. The propagation to the right of the barrier is the same as if the opening was a point source

opening propagates the incident wave energy as a point source. Circular concentric waves are seen beyond the opening. This can be understood physically because the wave motion at the opening is indistinguishable from the oscillation of a point source itself. The re-propagation at the point source leads to the apparent bending of the incident wave around the edges of the barrier. This "bending" is called *diffraction*. Diffraction is strictly a wave phenomenon and can be treated in the same way as interference. There is no need for new concepts. The geometrical method developed by Christian Huygens in 1678 is a very useful abstraction for treating complicated problems in wave propagation. In this method, every point in a wavefront under consideration is treated as if it is a point source. All of the point sources are assumed to radiate, and the net new wavefront results from the interference of each of these point sources. The method can be applied to describe the wave propagation for plane and spherical waves. Moreover the reflection and refraction of waves can be derived by the Huygen's construction principle.

Huygen's construction method naturally leads to considering the interaction of waves with small point sources though only in a medium. In the case of electromagnetic waves which have oscillating electric and magnetic vectors, these oscillations induce the charge distributions in the atoms of a substance to re-radiate the energy. This re-radiation is generally called *scattering*, and the scattering pattern depends on the nature and organization of the point radiators, the atoms in a material. We can treat a material as containing a large number of regularly spaced centers separated by a distance much smaller than the wavelength of the incident wave. For example, a transparent crystal with centers at 0.5 nm is illuminated by visible light with a wavelength of about 500 nm. The wave will strike the centers and be re-radiated so that the phase differences from nearby centers cancel, and the wave that is seen as scattered back obeys the law of reflection. Thus scattering plus interference is a microscopic interpretation of reflection. Refraction is explained by considering the scattering of a wave forward, in the same direction as the incident wave. The scattering causes a series of repropagated waves whose phase differences interfere, leaving a resultant wave whose phase lags behind that of the unscattered wave. This has the effect of shortening the distance that the incoming wave penetrates into the medium in an interval of time; hence, in comparison, the wave has a smaller velocity in the medium.

The principle underlying x-ray diffraction studies (and diffraction studies by matter waves such as neutrons and electrons) can also be understood in terms of Huygen's principle. Then the wavelength is comparable in dimension to the spacing between centers of scattering. There will be appreciable differences in the phase between the waves scattered by neighboring centers. For a single row of scattering centers, the scattered waves will be in phase if Δr_1 and Δr_2 are equal (Fig. 29.6a). This condition will occur when the angle of scattering is equal to the angle of incidence of the wave. If a second row of scattering centers is now considered (Fig. 29.6b), the scattering angle will still be equal to the incident wave.

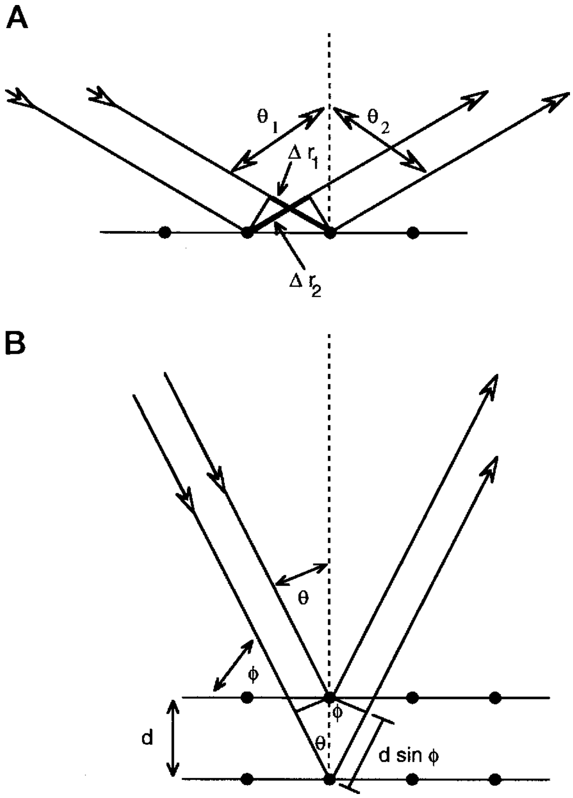


Fig. 29.6 Geometry of the Bragg scattering condition

However, only under specific conditions will the scattered waves from the two layers be in phase. This condition is met when the path difference for the two waves is $2d\sin\phi$, where ϕ is the complement of the angle of incidence: $\phi = 90° - \theta$. When the path difference is an integral number of wavelengths, the scattered waves will be in phase from all layers in a sample with regularly spaced centers. This condition is called the *Bragg scattering condition* and was described in 1913 by Sir William Lawrence Bragg. A strongly scattered wave will be seen when the angle satisfies the Bragg equation:

$$2d\sin\phi = m\lambda \tag{29.7}$$

At other angles destructive interference from the scattering centers will cancel out the scattering wave. These scattering patterns are the basis for determining the structure of molecules by x-ray diffraction methods.

We are able to learn a great deal about the structure of materials because of the scattering of light waves. This is because light is composed of oscillating electrical and magnetic waves, and atoms and molecules interact at the chemical level largely because of their electrical charge distributions. When the oscillating electromagnetic fields interact with the atoms, the atoms become radiators of the light and act as scattering centers by re-radiating the light. Scattering phenomena such as polarization, birefringence, optical rotatory dispersion, and circular dichroism can all be treated with the wave mechanics that we have discussed to this point.

29.3 X-Ray Diffraction Is a Powerful Fool for Structure Determination

We are now able to understand the application of scattering to solving the problem of crystal structure. X-ray diffraction studies remain among the most powerful and important tools of the modern biophysical chemist. We cannot attempt a detailed or practical treatment of the subject here but will survey the topic and references for further investigation by the student are provided in the reading list.

Consider the one-dimensional problem of the diffraction pattern generated by the scattering of light from a long line of regularly spaced centers. When illuminated, these centers will scatter photons in all directions but reinforcement of the scattered rays will be seen in limited directions. The reinforcement will be visible when the path distance between the scattered rays is equal to an integral multiple (l) of the wavelength (λ) of the illuminating light. We arrange the illuminating light so that it falls on the scattering centers at a perpendicular angle. The angle between the scattered light and the row of scattering centers is (γ). Thus reinforcement occurs under the condition:

$$l\lambda = c\cos\gamma \tag{29.8}$$

where c is the spacing between the centers. If the line of scatterers is sufficiently long, all wavefronts except those described by Eq. (29.8) will be cancelled out and only cones of scattered radiation remain. These cones can be projected onto a screen that is itself normal to the illuminating light. The projected diffraction lines will be shaped like hyperbole, and their spacing will be nearly proportional to $\cos \gamma$. Thus the line spacings will be inversely proportional to the spacing of the centers; close spacing of the scattering centers leads to wide spacing of the diffraction pattern. We can write Eq. (29.8) more generally so that the angle of illuminating light can be any angle off the perpendicular (γ_0):

$$l\lambda = c\,(\cos \gamma - \cos \gamma_0) \tag{29.9}$$

Now we consider the case of a two-dimensional array, a grid of scattering centers with spacing a and d on the x- and z-axes, respectively. The grid is aligned in the x–z axis with illumination perpendicular to it. Two equations are written that require simultaneous solution, Eq. (29.9) and:

$$\frac{h}{m\lambda} = a\,(\cos \alpha - \cos \alpha_0) \tag{29.10}$$

Constructive interference will occur only at the intersections, the conical sections scattering from the x- and z-planes and thus spots of scattered light will be captured on a photographic plate placed parallel to the grid.

We are generally interested in the use of x-ray diffraction to define the three-dimensional crystal structure of a molecule. If we consider a crystal with a unit cell with each angle at 90° (an orthorhombic crystal) and spacings of a, b, and c, we add a third equation to our previous two. These three equations are called the *von Laue equations*:

$$\frac{h}{m\lambda} = a\,(\cos \alpha - \cos \alpha_0) \tag{29.10}$$

$$k\lambda = b\,(\cos \beta - \cos \beta_0) \tag{29.11}$$

$$l\lambda = c\,(\cos \gamma - \cos \gamma_0) \tag{29.9}$$

One physical consequence of the required simultaneous solution of the von Laue equations is that arbitrary orientations of the crystal to the illuminating beam will not lead to a general solution. Only particular orientations will generate a diffraction pattern. These orientations can be found by methodically rotating the crystal and observing the appearance of diffraction patterns. The relationship of the Bragg law (Eq. 29.7) to the von Laue equations can be seen by solving the question: Given arbitrary orientation, will there be constructive interference at an angle 2θ with respect to the beam? Without showing the derivation we can use some algebra to generalize the Bragg equation to our orthorhombic crystal:

$$\lambda \left(\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2} \right)^{1/2} = 2 \sin \theta \qquad (29.12)$$

What happens in an x-ray scattering experiment? X-ray photons interact with electrons and are scattered. If the incident and the scattered wavelength are of the same wavelength then the scattering is elastic. Elastic scattering will generate diffraction patterns because of interference. Though we talk about scattering centers as point sources, the electrons that actually cause the scattering are characterized as distributions of electron density around a point in a molecule. The greater the electron density, the larger the amplitude of the scattered light. The relative intensity of the scattered ray (f_0) is a function of sin $\left(^{\theta}/_{\lambda}\right)$ (Fig. 29.7). Thus substitution of heavier atoms in a molecule such as a protein can be used to influence the scattering patterns and this will have practical value in solving the diffraction patterns generated by macromolecular scattering.

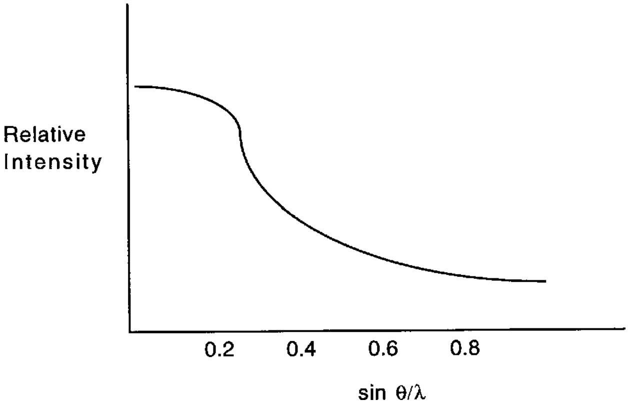


Fig. 29.7 Shape of the curves for relative scattering intensity with respect to sin $\left(\frac{\theta}{\lambda}\right)$. The ratio of the scattering intensity for an electronic distribution of an atom to the scattering by a point electron is f_0. This graph represents the data for a carbon atom. As $\theta \rightarrow 0$, f_0 approaches the number of electrons per atom. At this angle, all of the scattering is forward and there is 100% constructive interference

We have seen that the spacing of scattering centers will lead to patterns of constructive interference. The distance between the interference lines or dots provides information with respect to the distance of separation of the centers. In addition, each interference spot will have varying intensity depending on the electron density of the scattering center. Both the position and intensity of the spots comprising the diffraction pattern carry the information about the molecular crystal's structure.

A crystal is an orderly array of atoms or molecules that are arranged in a three-dimensional lattice. In Chapter 3 we spoke of the ideas of symmetry; the concept of translational symmetry is important when considering the array of atoms in crystals. Any crystal can be conceived as being formed by a repeating fundamental structural unit. This fundamental unit or *unit cell* is translated in coordinate space along the axes of the lattice making up the crystal. Because all of these unit cells are equivalent except for the coordinate transformation by spatial translation, the unit cell is basic in the description of the crystal's symmetry. Usually a unit cell is chosen that has the highest order of symmetry and is also of the smallest possible volume. The types of unit cells and their symmetries are listed in Table 29.1. The positions of the atoms in a unit cell can vary between the corners, faces, and body without destroying the unit's symmetry. These variations give rise to the *Bravais lattices* (Fig. 29.8). If a unit cell has atoms in only the corners it is called *primitive*; *body centered* if there is a point at its center; *side centered* if both the center of two opposite faces and the corners are occupied; and *face centered* if the corners and centers of all six faces are occupied.

Table 29.1 Unit cells and the symmetries associated with them

Unit cell	Geometry axes	angles	Essential symmetry	Example
Triclinic	a; b; c	α; β; γ	None	$K_2Cr_2O_7$
Monoclinic	a; b; c	$\alpha{=}\gamma{=}90°$; β	1 twofold axis	Monoclinic sulfur
Orthorhombic	a; b; c	$\alpha{=}\beta{=}\gamma{=}90°$	3 perpendicular twofold axes	Rhombic sulfur
Rhombohedral	$a{=}b{=}c$	$\alpha{=}\beta{=}\gamma$	1 threefold axis	Calcite
Tetragonal	$a{=}b$; c	$\alpha{=}\beta{=}\gamma{=}90°$	1 fourfold axis	White tin
Hexagonal	$a{=}b$; c	$\alpha{=}\beta{=}90°$; $\gamma{=}120°$	1 sixfold axis	Graphite
Cubic	$a{=}b{=}c$	$\alpha{=}\beta{=}\gamma{=}90°$	4 threefold axes in a tetrahedral arrangement	Rock salt

Unit cells are classified into one of the seven listed categories based on the symmetry that each possesses around certain axes

The von Laue or Bragg relations provide information about the spacing of the points in the crystal lattice. It will be the diffraction of the incident x-ray by a crystal plane that will give rise to an analyzable diffraction pattern. Thus it is valuable to have a nomenclature for identifying the crystal planes. The *Miller indices* are such a system. It is usually easier to visualize a two-dimensional lattice when first gaining experience with naming crystal planes and we will do so here. A two-dimensional lattice is drawn in Fig. 29.9. Each point in the lattice is characterized by a set of numbers on the a- and b-axes. A set of parallel lines can be drawn. We chose the heavy line (line #1) and starting at the origin ($0a$, $0b$), then write the coordinates at which it intercepts the two axes. These are ($4a$, $4b$). Now we take the reciprocals of the coefficients ($\frac{1}{4}$, $\frac{1}{4}$). The Miller indices are found by finding the smallest number that can be multiplied times the reciprocals to give integral indices, in this case 4. Thus the Miller indices are (1, 1). The process for lines 2, 3, and 4 are listed in Table 29.2. The general Miller index is written in terms of coinciding with the a-, b-, and c-axes. The smaller the value of h, k, or l, the more nearly parallel to the plane

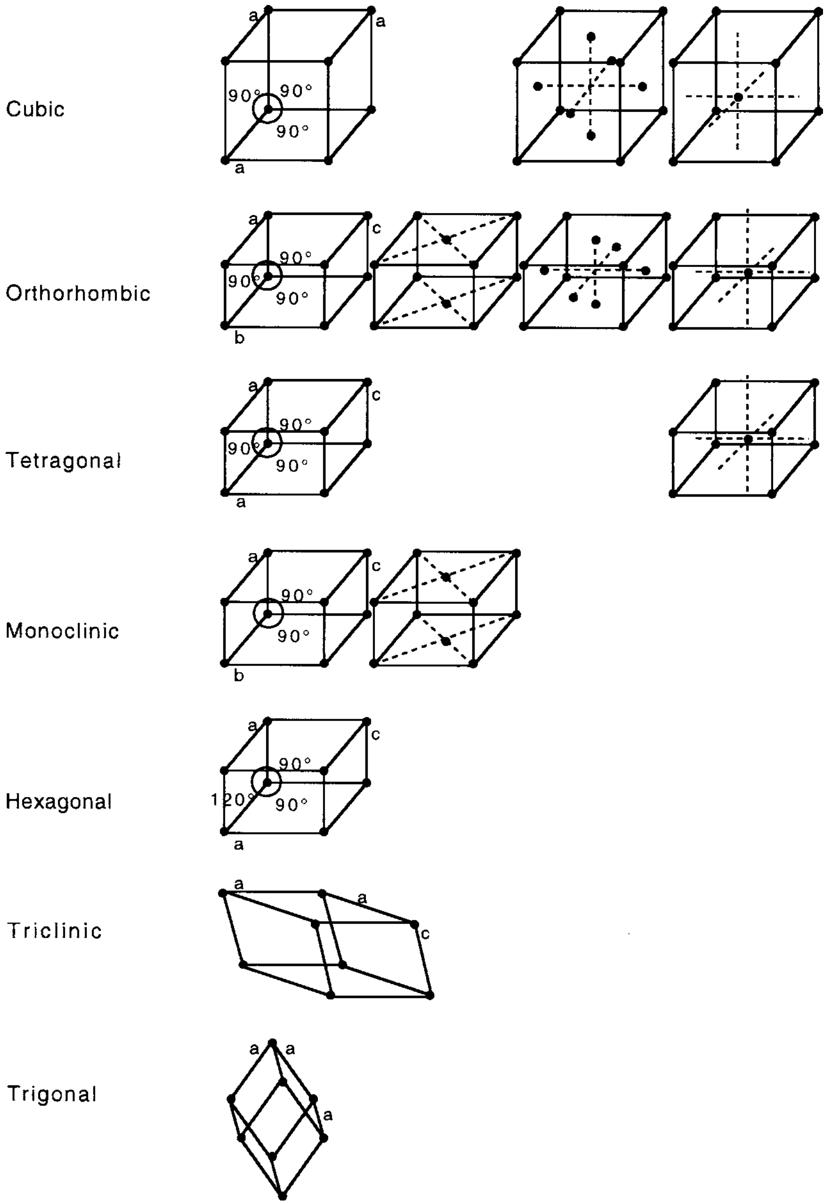


Fig. 29.8 Unit cell geometries and the Bravais lattices (from Giacorazo 1992, by permission of Oxford University Press.)

of the axis *a*, *b*, or *c*, respectively. Thus a Miller index of 0*kl* is parallel to the *a*-axis, *h*0 l is parallel to *b*, and so on. The Miller indices are related to the coefficients *hk* and *l* in the von Laue equations. This can be appreciated by recognizing that the

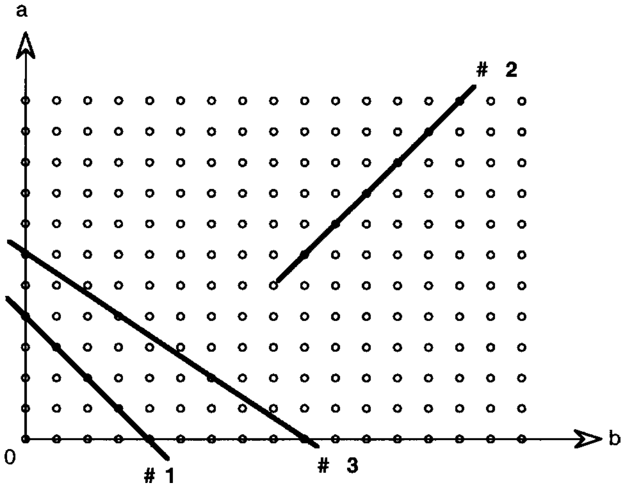


Fig. 29.9 Set of two-dimensional lattices for determining Miller indices

Table 29.2 Calculation of Miller indices

	Intercepts (*a*, *b*, *c*)	Reciprocals	Miller indices
Line # 1	4, 4, ∞	$\frac{1}{4}, \frac{1}{4}, 0$	1, 1, 0
Line # 2	–3, 3, ∞	$-\frac{1}{3}, \frac{1}{3}, 0$	$\bar{1}, 1, 0$
Line # 3	6, 8, 0	$\frac{1}{6}, \frac{1}{8}, 0$	4, 3, 0

A non-intersecting plane has a coefficient of ∞ which becomes 0 when the reciprocal is taken. Negative indices are written with a bar over the number. Usually we are interested in three-dimensional axes and write these as a Miller index (*hkl*) for the axes *abc*

separation of planes in a rectangular array (orthorhombic) built from unit cells with sides *a*, *b*, and *c* is

$$\frac{1}{d^2} = \left(\frac{h^2}{a^2} + \frac{k^2}{b^2} + \frac{l^2}{c^2}\right) \tag{29.13}$$

Compare this expression to Eq. (29.12). The fact that a relationship between the spacing of the scattering planes, the Miller index, the Bragg relation, and the von Laue equations exists should be obvious. Details and mathematical proof of the relationship can be found in the references at the end of the chapter.

We continue with consideration of an orthorhombic cell as the repeating unit in a crystal because it is relatively straightforward. The following analysis can be applied to other geometries in which the axes are not all at right angles to one another. In

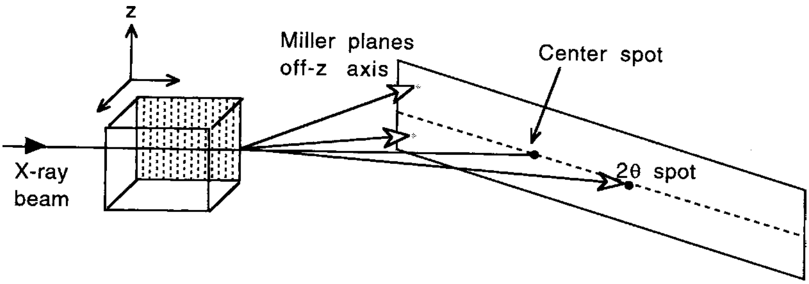


Fig. 29.10 The diffraction pattern generated by a crystal plane will be (**a**) along the equator in this 010 plane and (**b**) off the equator in other planes

these cases the mathematical expressions become substantially more complicated. If we have a single crystal composed of our orthorhombic unit cell, we can rotate the crystal and hence the unit cell. If a monochromatic beam of x-rays is passed into a single crystal that can be rotated a pattern can be seen on a screen. The crystal is oriented as shown in Fig. 29.10 with the *z*-axis perpendicular to the beam. When a 010 crystal plane makes the Bragg angle θ with respect to the beam, a diffraction spot will be produced on the equator of the crystal image at an angle 2θ. This equatorial line will also have other spots that correspond to the higher orders of diffraction ($h =$ 2, 3, etc). Finally, spots will be found for planes of higher Miller index orders (020, 030, etc). The reciprocal-space nature of the diffraction pattern will determine the spacing of the spots: the larger the spacing in the crystal, the closer the diffraction spot will appear to the center of the pattern. Miller index planes that are not parallel to the *z*-axis will also produced diffraction spots but these will be off the equatorial line. An example of a single-crystal diffraction pattern is shown in Fig. 29.11.

An alternative to producing a diffraction pattern by rotating a crystal in a monochromatic beam exists. In the powder method, a monochromatic beam illuminates a powdered sample of the material under study. In a powder some crystallite will be found that satisfies the Bragg relation without having to methodically rotate the sample. The diffraction pattern generated is called a *powder spectrum* and is identified by the appearance of concentric rings in contrast to the spots generated from a rotating crystal method.

We have discussed how a diffraction pattern can be generated from a lattice of point scatterers. However, our interest in biology is usually to do just the reverse: use a diffraction pattern to learn about the molecular structure of a compound. This problem is complicated by several issues:

1) Though the intensity of the scattering from a point scatterer is independent of the angle of scattering, this is not true for atoms which are scatterers characterized by a diminishing electron density with increasing radius off-center.
2) The bright spots in a diffraction pattern are generated by constructive interference, but the intensity of the spot will be equal regardless of whether the peaks or troughs of the waves constructively interfere. Thus we cannot know the phase

of the wave making the spot by looking at the pattern. This means that when we reconstruct the lattice from the image, we may be out of focus without any way of knowing it.

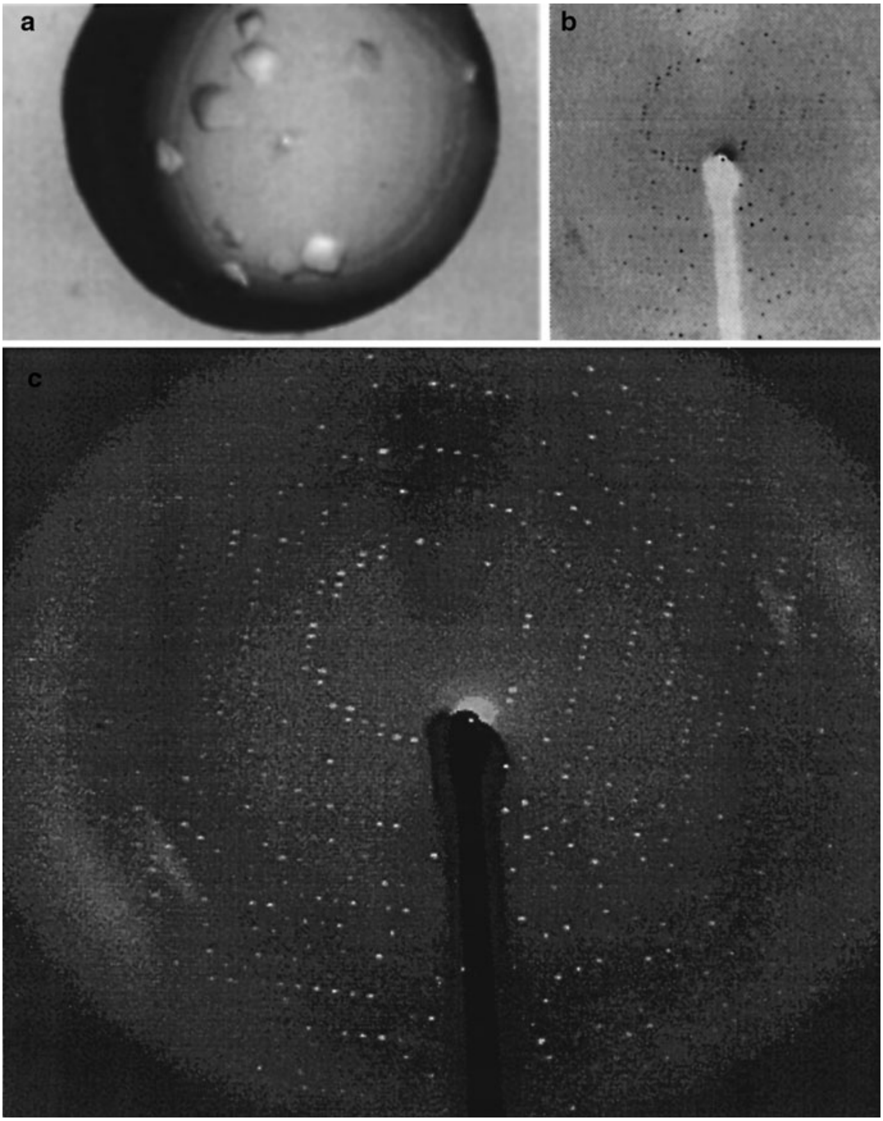


Fig. 29.11 Single-crystal x-ray diffraction pattern of β_2-glycoprotein 1. The crystals seen in the *upper left panel* were grown by vapor diffusion in hanging drops. The crystals were then diffracted with the resulting pattern (2° oscillation). The *lower panel* is an image enhancement of the smaller diffraction pattern and allows the *dot pattern* consistent with an orthorhombic crystal to be appreciated (Courtesy Drs. Bin Lu and Mary T. Walsh)

3) In a three-dimensional lattice, there are potentially many scattering planes that could contribute to the diffraction pattern. We cannot be sure of the location of these planes unless we know the phase of the scattered waves that are interfering to make the pattern.
4) Finally, there are different atoms making up the unit cells in complex molecules such as proteins. Thus the phase problem, the scattering intensity problem and the scattering plane problem all can have another level of interference that modifies both intensity and spot pattern.

Imagine a simple lattice composed of two scattering planes both of which satisfy the Bragg condition and therefore scatter waves in phase for their plane. However, each of these planes is offset from the other by distance x. Thus though each plane is in phase with itself, the scattered waves will be out of phase with one another. The phase difference is

$$\Delta\phi = 2\pi \frac{hx}{a} \tag{29.14}$$

where a is the separation between the centers of the congruent planes and h represents the order of diffraction. Choosing one atom as the origin of the unit cell we can write for any set of planes:

$$\Delta\phi = 2\pi \left(\frac{hx}{a} + \frac{kx}{b} + \frac{lx}{c} \right) \tag{29.15}$$

We are interested in summing all of the waves coming from a group of scatterers (found in a unit cell). Looking ahead, this problem means that we will be summing a wide variety of amplitudes coming from centers of diverse scattering power. Therefore to keep the mathematics as simple as possible the analysis is usually performed using complex numbers rather than using all sin and cos notation. We write either

$$f \cos\left(2\pi \nu t + \Delta\phi\right) \quad \text{or} \quad fe^{i(2\pi \nu t + \Delta\phi)} \tag{29.16}$$

The rays scattered by a point in a unit cell will be given by $f_j e^{i\Delta\phi}$, where f_j is the atomic scattering factor and $\Delta\phi$ is the phase. This expression is for those waves scattered at the Bragg angle. The amplitude of the sum of the scattered waves in a unit cell of n atoms is given by

$$F\left(h,k,l\right) = \sum_{j=1}^{n} f_j e^{i\Delta\phi} \tag{29.17}$$

This term is called a *structure factor*. It should be apparent that each reflection seen in a diffraction pattern can be composed of hundreds to thousands of components from the atoms in a unit cell of a biological crystal. Extracting the structural information from such a pattern may seen hopeless.

However, a solution to this problem exists. The structure factor could be written explicitly in terms of the coordinates of the scattering centers (*Xj*, *Yj*, and *Zj*):

$$F(h,k,l) = \sum_{j=1}^{n} f_j e^{2\pi i(hX_j+kY_j+lZ_j)} \tag{29.18}$$

The scattering amplitude contains via Fourier transformation, the scattering or electron density at each point $\rho(X,Y,Z)$ in the unit cell.

$$\rho(X,Y,Z) = \frac{1}{V}\sum_h \sum_k \sum_l F(h,\ k,\ l)e^{-2\pi i(hX+kY+lZ)} \tag{29.19}$$

V is the volume of the unit cell. Thus, measuring the scattering amplitude, performing a Fourier transform, and finding an electron density map for the unit cell will provide the molecular structure. We are assuming a knowledge of phase, and we do not have easy access to that piece of the puzzle. The reason is that we measure intensity, not amplitude. As we already learned, intensity is the square of amplitude. When we square the amplitude we lose the sign of the amplitude (i.e., the phase). Thus we can know the magnitude but not the direction of the interference wave.

A solution exists. If a heavy atom that will scatter strongly is added to the crystal under study and we compare the native molecule with the one that is isomorphously replaced, we can deduce the phase of each reflection. This is possible because the strong scatterer can be located and give a strong phase interference to all of the reflections. Assume that it gives an interference wave of positive phase. The already positive reflections will become brighter and the negative phase reflections will dim. By comparing the difference between the two diffraction patterns, a *Patterson* difference map can be drawn that gives us the phase information that we seek to solve the crystal structure. In practice a double isomorphous substitution must be performed in order to know with confidence the sign of the first substitution's phase angle.

29.4 Scattering of Light Rather Than Its Absorption Can Be Used to Probe Molecular Structure and Interaction

29.4.1 Rayleigh Scattering

Light is not always absorbed when it interacts with matter. Under some conditions the electric field of a photon transiently disturbs the charge distribution of the molecule but no transition to another energy state occurs; the photon is re-radiated at the same wavelength as the incident photon. Because the photon does not induce a change from one state to another there are no selection rules governing the interaction. The incident and emitted photons are the same energy and the interaction is therefore elastic. Scattering of this kind is called *Rayleigh scattering*. The interaction may be visualized as follows. The electric field of the incident photon interacts with the charge distribution of any molecule. All molecules are polarizable and depending on the ease of molecular polarizability (α) the incident electric field will induce an oscillating dipole moment in the molecule. The magnitude of the induced dipole (μ_{distort}) depends on the electric field at the molecule and its polarizability:

$$\mu_{\text{distort}} = 4\pi\varepsilon_o\alpha E \tag{29.20}$$

but E is varying and so

$$\mu_{\text{distort}} = 4\pi\varepsilon_o\alpha E_o \cos 2\pi vt \tag{29.21}$$

If the molecule is isotropic, the induced dipole will vibrate in the same direction as the electric vector. An anisotropic molecule will couple to the incident field and re-radiate depending on the resultant shape of the distorted induced dipole. The oscillating dipole will become a radiator of an electromagnetic field and will propagate a wave both forward and backward relative to the vector of the incident wave. Thus the wave is scattered. The issue of importance is the intensity of the scattered wave with respect to the radiating dipole. The radiated electric field at a distance r from the dipole and at an angle ϕ with respect to the axis of polarization can be determined from electromagnetic theory. The intensity of the scattered wave (i) is the square of its amplitude and if related to the intensity of the incident beam (i_0) can be shown have the form of

$$\frac{i}{i_o} = \frac{16\pi^4\alpha^2 \sin^2\phi}{r^2\lambda^4} \tag{29.22}$$

(This applies to an isotropic particle irradiated by polarized radiation but an equation of similar form can be obtained for unpolarized incident light.) The intensity of the observed scattered wave decreases with r^{-2} and is dependent on ϕ. The intensity increases strongly with shorter wavelengths of light. Scattering is maximal in the same direction as the incident wave and zero in the direction of the dipole vibration. We see the sky because of the scattering of light by the air molecules and certain contaminants. It has a blue color because blue light is more strongly scattered and thus dominates the color spectra of what we see.

We have described the situation for a single scatterer. What happens when light is incident on a condensed phase such as a crystal or solution? In these cases the scattered waves will interfere with one another to varying degrees and the scattering intensity will be increased or decreased depending on the constructive and destructive interference of the waves. Dissolved macromolecules scatter quite strongly and the intensity and angle of the scattered light can provide information about the absolute molecular weight and to some degree the shape of the molecules. A more detailed treatment of this application of scattering can be found in the references at the end of the chapter.

29.4.2 Raman Scattering

We have only described elastic scattering to this point. The re-radiation of the incident light at the same wavelength is true only when the polarizability of the molecule is constant with respect to the incident coupling. This is not necessarily true. In vibrating molecules the polarizability will vary depending on the inter-nuclear distances between the vibrating atoms. Our treatment of vibrating molecules takes the form of a harmonic oscillator and the polarizability with respect to this motion $\left[\alpha\left(v'\right)\right]$ will be

$$\left[\alpha\left(v'\right)\right] = \alpha + \beta\cos 2\pi v't \tag{29.23}$$

The induced dipole will be

$$\begin{aligned}\mu_{\text{distort}} &= 4\pi\varepsilon_o\alpha\left(v'\right)E \\ &= 4\pi\varepsilon_o\left(\alpha E_o\cos 2\pi vt + \beta E_o\cos 2\pi vt\cos 2\pi v't\right)\end{aligned} \tag{29.24}$$

The first term is recognizable as the Rayleigh scattering term with elastic radiation. The second term represents scattering above and below the wavelength of the incident light by a factor of ($v \pm v'$). This inelastic scattering is called *Raman scattering* and reflects the vibrational modes of a molecule. Thus Raman spectra can provide information about the vibrational motions of a molecule, i.e., in the absorption region associated with infrared spectra. The Raman spectra is generated because there is an interaction between the incident light and the molecule that leads to energy transfer. The energy transferred to the molecule leaves it in a higher vibrational energy level (v – v') but energy can be transferred out of the molecule if a vibrationally excited state is induced to scatter (v + v'). Raman scattering has a selection rule: the polarizability of the molecule must change with vibration. This is a different selection rule than IR spectroscopy thus Raman scattering often yields information about different transitions than infrared adsorption spectroscopy and therefore is an adjunctive methodology. The Raman bands are very faint compared to the Rayleigh scattering intensities because $a >> b$ by a factor of 10^4 and most modern Raman instruments use lasers to generate intense and highly collimated light sources. A final point about Raman scattering is that the scattering spectra can be found at any wavelength: thus the wavelength of the illuminating source can be optimized away from solvent adsorption bands or to the ideal response parameters of the measuring instrumentation. A Raman spectra can also be seen at the wavelength that an electronic absorption transition takes place. This phenomena is called *resonance Raman scattering* and has the attraction that the Raman spectra is up 10^3 times more intense than normal Raman scattering.

29.4.3 Circular Dichroism and Optical Rotation

Most biological molecules are asymmetric, either because they contain asymmetric carbon atoms or because the supramolecular structure such as a helix winds either in a right- or left-handed fashion. These asymmetries are manifest in electronic charge distributions that will interact differentially with the electric vector of light. If we consider the interaction of light as a wave that is scattered by these asymmetric centers of charge, we can appreciate that the direction and intensity of the scattered light will reflect the internal arrangements of these molecules. We are generally concerned with the interactions of plane or circularly polarized light as it interacts with anisotropic molecules. Materials may be *birefringent* or *dichroic*. A material that is birefringent has a refractive index that differs depending on the plane of incidence of a polarized beam of light. Such materials will rotate a polarized beam differently because of this difference in refractive index. Alternatively stated, the velocities of propagation for a right- or left-polarized beam of light will be different. This difference can be measured with a polarizer, which can quantitate the rotation of the beam of light. This method can be used to explore the amount of coil, helix, and β-sheet in a macromolecule.

Chiral molecules will absorb right- and left-polarized light in a differential fashion. Thus for an asymmetric chromophore:

$$\Delta A = A_L - A_R \tag{29.25}$$

and, if a sample with circular dichroic properties is illuminated with an incident beam of plane polarized light, there will be both a phase shift due to the refractive index difference (the birefringent effect) and a difference in the absorbance of the right-and left-circularized components of the incident beam. The net result will be rotation, but with a difference between the left and right vectors yielding an elliptically polarized beam. The ellipticity can be either positive or negative and is measured in terms of the molar ellipticity:

$$\theta\ (\text{rad/cm}) = \frac{2.303\,(A_L - A_R)}{4\,l} \tag{29.26}$$

where l is the path length of the sample. CD studies are especially useful because the amount of secondary structure in a macromolecule can be related to the molar ellipticity in the far-UV absorbance range. Thus CD measurements under differing conditions can report on the secondary structure changes in the macromolecule (Fig. 29.12).

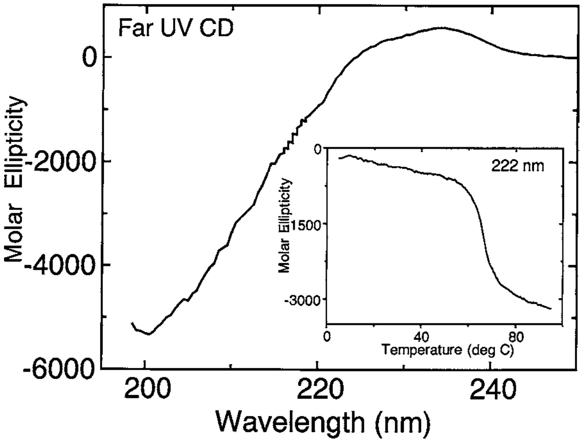


Fig. 29.12 Circular dichroism can be used to monitor secondary structure during unfolding experiments as in this example of the thermal unfolding of β_2 glycoprotein I in aqueous buffer. Following the molar ellipticity at 222 nm shows a highly cooperative unfolding with a *T*m of 62°C (Courtesy of Drs. Bin Lu and Mary T. Walsh)

Further Reading

General

Feynman R.P., Leighton R.B., and Sands M. (1963) *The Feynman Lectures on Physics*, Volume 1. Addison-Wesley, Reading, MA.

Freifelder D.M. (1982) *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*, 2nd edition. W.H. Freeman, New York.

Nölting B. (2006) *Methods in Modern Biophysics*, 2nd edition. Springer-Verlag, Berlin.

Serdyuk I.N., Zaccai N.R., and Zaccai J. (2007) *Methods in Molecular Biophysics*. Cambridge University Press, Cambridge.

Sheehan D. (2009) *Physical Biochemistry: Principles and Applications*, 2nd edition. Wiley-Blackwell, Oxford.

Van Holde K.E., Johnson W.C., and Ho P.S. (2005) *Principles of Physical Biochemistry*, 2nd edition. Prentice Hall, Upper Saddle River, NJ.

History

Farmelo G. (1995) *The discovery of X-rays*, *Sci. Am.*, **273**, **5**:86–91.

Crystallography

Cantor C.R. and Schimmel P.R. (1980) *Biophysical Chemistry, Volume II*. W.H. Freeman, New York.

Giacovazzo C. (ed.), Monaco H.L., Viterbo D., Scordari F., Gilli G., Zanotti G., and Catti M. (1992) *Fundamentals of Crystallography*. International Union of Crystallography, Oxford University Press, New York.

Lisensky G.C., Kelly T.F., Neu D.R., and Ellis A.B. (1991) The optical transform, *J. Chem. Educ.*, **68**:91–96. (A nice practical article to read along with the Perutz chapter, Diffraction Without Tears.)

Perutz, M. (1992) Appendix 1 – mathematical principles of x-ray analysis, In *Protein Structure: New Approaches to Disease and Therapy*, W.H. Freeman, New York, pp. 261–275.

Perutz M. (1992) Chapter 1 – Diffraction without Tears: a pictorial introduction to x-ray analysis of crystal structures, *In Protein Structure: New Approaches to Disease and Therapy*. W.H. Freeman, New York, pp. 1–39. (A non-mathematical treatment. Elegant and insightful.)

Scaterring Techniques for Evaluation of Macromolecules

Cantor C.R. and Schimmel P.R. (1980) *Biophysical Chemistry, Volume II.* W.H. Freeman, New York.

Freifelder D.M. (1982) *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*, 2nd edition. W.H. Freeman, New York.

van Holde K.E. (1985) *Physical Biochemistry*, 2nd edition. Academic, New York.

Raman Spectroscopy

McCreery R.L. (1996) Analytical Raman spectroscopy: an emerging technology for practical applications, *Am. Lab.*, **Feb**:34X–34JJ.

Pelletier M. and Davis K. (1996) Raman spectroscopy: the next generation, *Am. Lab.*, **Feb**: 34C–34 N.

Chapter 30
Analysis of Structure – Microscopy

Contents

30.1 Seeing Is Believing

We know an object is located in state space when we can interact with it. We know that the choice of observables, the method of measuring them, and the abstractions used to connect the observables to a formal-natural system model are the essence of modern biophysical science. Furthermore, we are bound to a great degree by our own innately human system for observing, which is largely visual. Few methodologies in the biophysical sciences are more satisfying than those that employ *visualization* of the state space under consideration. The human brain is naturally organized to extract data from visual images. Thus *microscopy* and its related imaging methodologies are powerful tools for extracting and mapping information about a system. A word of caution is necessary. While the psychology of image and

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_30,

iconographic power as a natural consequence of the neurological functioning of the human brain is beyond our scope here, as human researchers we must be skeptical of the conclusions derived from our iconoscopic powers. Readers, teachers, and marketing departments are universally aware of the power of a good pictorial representation to drive home a point. But abstraction is taken to its highest level in pictures because the human brain is naturally driven to fill in empty data sets. Having done so, however, the brain usually losses the distinction between observed and abstracted. This neurological determinism is, in part, the biological explanation for many of the paradoxes in biophysical studies. It certainly underlies the unfortunate Aristotelian dogma that the only "true" knowledge was to be gained through the observations made by the primary human senses.

Visualization often means the generation of a *pseudoimage,* which is a graphical representation of a function or a set of variables by a computer output device. In many cases these images are wholly abstractions given form by the computer or the pen and assigned authority by the mind's eye. However, the science of visualizing an object on a small scale by the interaction of light with one or several of the object's properties is the essence of microscopy. Traditional microscopy is an application of optics in which lenses are used to focus light from an object on the observer's eye. Many of the limitations of microscopic analysis reflect the wave–particle duality. Microscopy in many ways is the prototype physical measurement because it allows a specific interaction to be related point to point in a coordinate system (usually three-dimensional Cartesian system) thus naturally generating a "picture" of the object. Modern microscopy still includes visible light microscopy but a wide variety of other interactions are now studied on a three-dimensional point-to-point basis thus allowing us to "see" the physical coordinates or locations of topography, electrical field, electron density, chemical properties, etc.

With the caveat of the preceding paragraphs, one of the most compelling and important advantages of microscopy in biophysical investigations is that we generally focus our attention on the cell and the physical dimensions in which we observe cell structure, organization, and components. Even the powerful newer techniques including near field microscopy and scanning force microscopy essentially are bounded by a state space that is cellular in magnitude. This perspective allows the biophysical chemist to appreciate and often achieve a sophisticated degree of quantification that relates the physical form and coordinates of chemical species with respect to the cellular form and organization. As an example, appreciation of the chemistry and enzymology of the glycolytic pathway, the Krebs cycle, and the electron transport chain can be gained without knowing that there is physical compartmentalization of these systems between the cytosol and the mitochondria. However, if the assumption were made that these reactions took place in a single phase system, it would be difficult to understand the chemistry of the carnitine acyltransferaces, which shuttle acetyl-CoA between the outer and inner mitochondrial compartments. In fact, there are diseases caused solely by defects in these transport systems when the metabolic enzymes are otherwise normal. Thus, without knowledge of the compartmentalization, the pathochemistry of diabetic ketoacidosis and the epilepsy associated with carnitine deficiency would be difficult to appreciate.

We will examine the variety of microscopic studies available to the modern biophysical scientist in this chapter. All are bound by several fundamental issues: resolving power, magnification, sample preparation, artifact identification, sample stability.

30.2 The Light Microscope Allows Visualization of Structures on the Dimensional Scale of the Wavelength of a Photon

In Appendix G, the geometric optics of forming an image with a lens is described. We have seen that under certain conditions the wave nature of light was dominant and light could be treated as a ray that travels in a straight path. These rays can be bent at the interfaces between two transparent objects. This forms the basis of geometric optics. Alternatively, we can emphasize the electromagnetic nature of light in which diffraction and interference are recognized. This underlies the ideas of physical optics. From a practical standpoint, these approaches are simply two sides of the same coin; the geometric approach is useful in understanding the ideas of image focus and aberration, while the physical approach leads more easily to an explanation of image contrast and the limits of resolution.

Let us consider image formation in a modern compound microscope. Start with an ideal simple lens having two convex curvatures each with a radius of curvature and focal point (called F and F'). An image is formed by a point-by-point translation of the *object*, which is said to be in the *object plane* into an *image* which exists in the *image plane*. Figure 30.1 illustrates this translation. Two light rays coming from the point of the arrow are represented so that one passes parallel to the lens axis and the other passes through the focal point. Both then pass into the lens. After passing through the lens, the light rays emerge and the first is refracted through the focal point *F* while the second now continues parallel to the lens axis. The point of the intersection of these two rays will form the point of the arrow in the image plane. The geometry of the light rays obeys the relation $aa' = ff'$ with the magnification

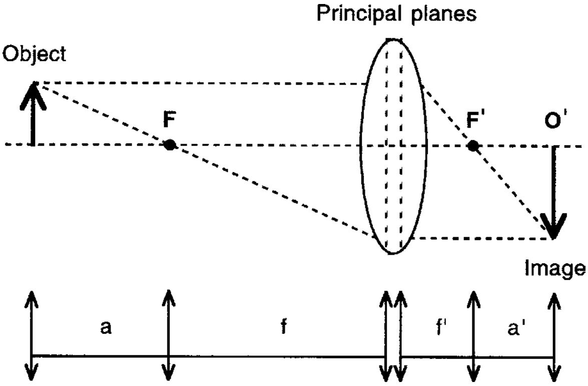


Fig. 30.1 Optical geometry of a simple lens

given by $-\frac{f}{a}$. The meaning of the negative sign is that the image is inverted in the image plane with respect to the object plane.

An ideal lens brings all the light rays from a single point in the object plane to a sharp focus in the image plane. However, real lenses fail to focus all the points in a single image plane and are said to have *aberration*. Commonly comes in several forms: chromatic, point-imaging, and astigmatic. *Chromatic aberration* occurs because the index of refraction is dependent on wavelength. Light of different wavelengths (white light) coming from a single point in the image plane will be focused into separate image planes. We use this effect to our advantage in prisms, but in a lens it makes the image fuzzy. Chromatic aberration can be corrected by constructing a lens system of various glasses each having different refractive indexes. Correctly engineered, the refractive index of the overall system can be made independent of wavelength. These lens systems are called *anachromatic*. In many applications, monochromatic light can be used and this also will minimize chromatic aberration. Most research grade microscopes use monochromatic light sources for this reason.

Point-imaging aberrations occur because monochromatic rays from a single point in the object plane will not necessarily pass through the same image point in the image plane. *Spherical aberration* occurs when this effect is caused by the refraction of a single object point by different parts of the lens with the consequence that the point in the image plane is not in focus. A related point-imaging aberration that leads to distortion of the shape of the object is called *coma* because this aberration gives a round point in the object plane a coma like appearance in the image plane thus destroying the symmetry of the image. Lenses constructed to eliminate point-imaging aberrations are called *aplanatic*.

Astigmatism is another symmetry destroying aberration in which the arrangement of object points in Cartesian space are warped so that the image does not maintain a linear point-to-point relationship with the object. This aberration usually becomes apparent only at high resolution where the information carried by each point becomes relatively more important for the overall observation of the object space. Lenses that are corrected for astigmatism are called *anastigmats*.

A simple microscope can be constructed with (1) a light source, (2) a method of holding an object to be viewed, and (3) a suitable lens that provides adequate magnification and correction of aberration. Unfortunately, in practice, a single lens microscope would require a lens with a very short focal length, and the observer's eye would have to be placed extremely close to the lens. To surmount these practical restrictions, a compound microscope is made by arranging an *objective lens* and a secondary lens system, the *eyepiece*, in succession (Fig. 30.2). This arrangement allows an object to be placed just beyond the focal point of the objective. The objective-magnified image is further magnified by the eyepiece and then focused onto the retina of the eye.

Though it is a convenient abstraction to treat the optical microscope using geometrical optics, it is more unifying to consider image formation as the result of the constructive and destructive interference of electromagnetic waves following

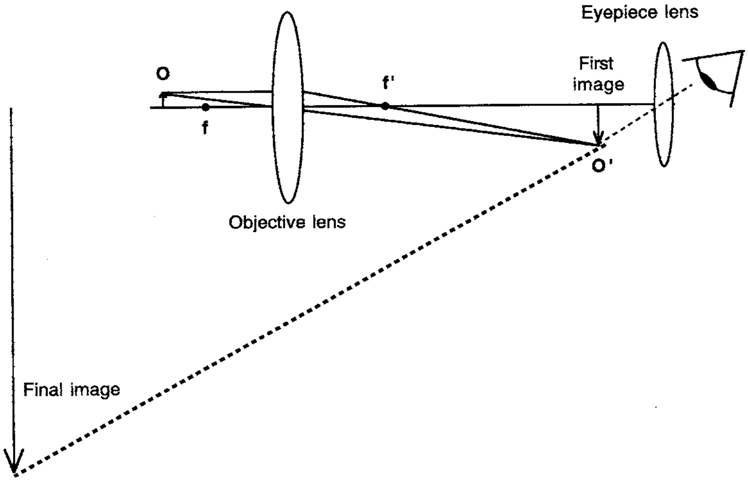


Fig. 30.2 Schematic of a compound microscope

diffraction. With this view all forms of image creation whether by light, UV, electron microscopes or x-ray/neutron diffraction studies are unified. Ernst Abbe's theory of microscopic vision is central to this discussion. Consider an object, an optical grating, illuminated by a light source of parallel rays. The light rays will be scattered by each point in the grating. The scattered rays will interfere constructively and destructively with each other, generating an interference pattern. This interference pattern can be collected and focused by the objective lens and can be seen in the back focal plane of the objective when the eyepiece is taken out of the line of sight. It is the light of the diffraction pattern that is collected by the eyepiece lens and again subjected to interference. This second interference pattern provides a true magnified image of the original grating. There is a reciprocal relationship between the object, its diffraction pattern, and its magnified image. In the microscope the image is reformed from the diffraction pattern because of the ability of lenses (optical and electron) to diffract those wavelengths. In x-ray and neutron diffraction systems, only the diffraction patterns can be seen because no suitable lens exists to perform the diffraction-image conversion with photons of such short wavelength. The process of translating this conversion is the technical concern of x-ray crystallography and has already been briefly discussed.

We are able to see the image of an object because light is scattered by the atoms of that object and the diffracted light waves then interfere to form a diffraction pattern. However, the lens also is a diffraction device, and the rays of light "refracted" in our geometrical optic approximation are actually being scattered by the atoms in the optical glass of the lens. Therefore the lens is actually an aperture, and the light passing through it will form diffraction patterns. In practice, a bright spot of light in the object plane will appear as a circle of light surrounded by a series of concentric

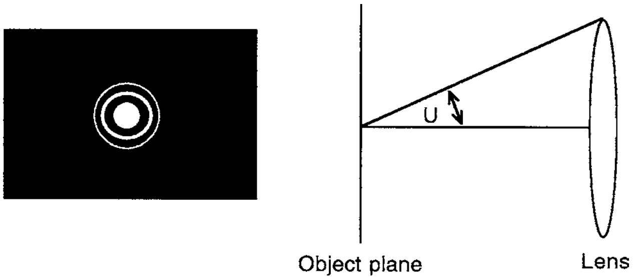


Fig. 30.3 (*left*) The airy disc is diffraction image of each point in the image field. Resolution depends on being able to separate these diffraction patterns from one another. (*Right*) The geometry used for determining the resolution of two points by a lens

rings, the result of interference. This pattern is called the *Airy disc*. The radius of the first dark ring encircling the central bright spot will be

$$\frac{0.6\lambda}{n}\sin U \tag{30.1}$$

with U being the angle made by drawing a line from the central axis of the lens to the edge of the lens (Fig. 30.3), λ the illuminating wavelength of light, and n the refractive index of the object side of the lens. If two object points are in close proximity, the ability of an observer to resolve them into two clearly distinct points, i.e., the resolution depends not on the separation of the two points but on the dimensions of their Airy discs. In fact, knowledge of the true resolution is limited by indeterminacy but the convention has been long adopted that Eq. (30.1) is the limit of resolution. Resolution can be improved by shortening the wavelength of the illuminating light (i.e., ultraviolet and electron microscopy), increasing the index of refraction on the objective lens side (placing the object under oil), and increasing U. U can be increased either by shortening the distance between the lens and the object or by increasing the diameter of the lens. The increase in resolution with higher U follows directly from our discussions of diffraction and the certainty of observed information. When the illuminating light is scattering by an object, the diffracted light leaves the object at many different angles, which leads to many different orders of diffraction. The state of each point in the object state space is given by the complete set of these many orders of diffraction. The only way to obtain complete information about the object's state space (image) is to collect all of the observable data which is carried in the totality of the diffracted orders. This ideal requires the largest possible collection angle between the object and the lens aperture, i.e. the largest U. A quantity called the *numerical aperture*, *NA*, is printed on objective lenses and is given by $n \sin U$. A lens with higher magnification that does not increase in numerical aperture provides "empty magnification" because the resulting image is larger but remains poorly resolved (Table 30.1).

Table 30.1 Magnification and NA in objective lenses

Magnification	Numerical aperture	Focal length	Working distance (mm)	Field diameter (mm)
10	0.25	16	5.50	2.00
20	0.54	8	1.40	1.00
40	0.65	4	0.60	0.50
40	0.95	4	0.25	0.20
95 (oil)	1.32	2	0.10	0.05

30.3 Visualization Requires Solving the Problem of Contrast

The ability to see the details of a system under microscopic observation requires *contrast* between point elements in the object's state space. Biological systems are generally transparent to light, and their visualization under bright field illumination is often nearly impossible. A variety of techniques have been developed that increase the contrast of cells thus improving visualization. Contrast enhancing techniques include staining of specimens histochemically with color reactions, and the microscopic techniques of dark-field, phase, polarization and interference microscopy. All of these techniques can be understood on the basis of our discussions of physical optics.

30.3.1 Dark Field Microscopy

In bright field studies, light coming from the condenser system illuminates the object plane. In the absence of a specimen in the object plane light uniformly enters the objective lens giving an evenly illuminated field of view. If an opaque disc is inserted into the condenser system an annular or hollow cone of light will fall on the objective plane. This circle of light can be adjusted so that the light surrounds the objective leaving the field dark (a dark field). Though the field is dark some light reaches this region and any object placed in the dark field will diffract some of this light. A number of the orders of diffraction will reach the objective lens and the object will appear as a very bright high-contrast object against a very dark field. The resolution of the object is quite poor because the information carried in the low orders of diffraction is lost, but the technique illustrates how diffraction can be used to dramatically enhance contrast.

30.3.2 Phase Microscopy

A common technique for the visualization of living cells and organelles is to convert the phase differences produced when light passes through organelles of differing refractive indices into intensity differences. The human eye is not sensitive to the phase differences of light striking it, although it is well adapted to intensity differences. We know that light passing through each point of a specimen in the object plane is diffracted twice, once by the specimen and again in the objective lens. Each point of the viewed image is the result of the combined interferences of each point of the object and each point of the objective lens. Though the image contains the information related to the varying diffractive indices of the specimen, the phase differences are generally too small to generate sufficient interference to be perceived as intensity differences. If the lens system is built to alter the phase of either the zero-order diffracted (undeviated) or greater than zero-order diffracted (deviated) light sufficiently prior to recombination in the image plane, enough destructive interference can be generated to create contrast enhancement. In the phase contrast microscope (Fig. 30.4) an annular light source is used just as in dark-field microscopy. This hollow cone of light passes through the object plane and falls as a ring of light on the objective lens. The image of the annular ring is thus formed in the back focal plane of the objective lens (eyepiece side). In order to induce the necessary phase shift, a *phase plate* is introduced at the back focal plane. This phase plate is an annular disc of material that slows the light so that the light traveling directly from the condenser annulus is now retarded one-quarter wavelength. (Alternatively, the wavelength could be advanced if the phase plate consists of a groove in the path of the annular ring). Any light that passes through a specimen placed in the objective plane and is diffracted (deviated) will not be focused by the objective onto the phase plate and thus will not have its phase altered. The deviated and the phase-enhanced-undeviated light will be recombined at the image plane where substantial interference will now occur. The interference patterns are seen by the observer as intensity differences, and contrast has been increased. The phase contrast method allows living cells to be observed in substantial detail and is widely used in cell culture and in cell biological applications. Other interference techniques have also been developed including Dyson, Hoffman, and Nomarski optics (Fig. 30.5).

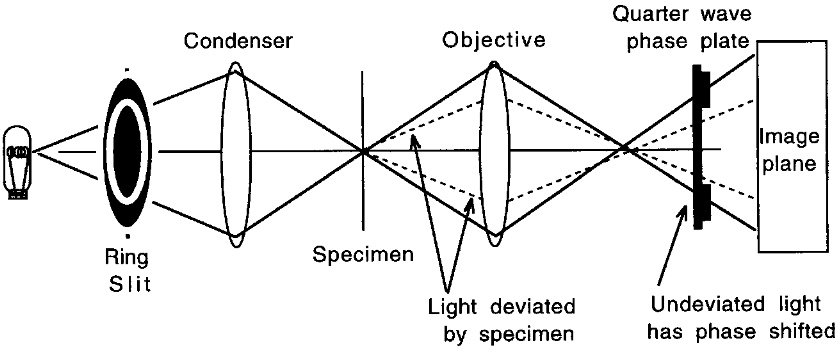


Fig. 30.4 Schematic of the optical train of a phase contrast microscope

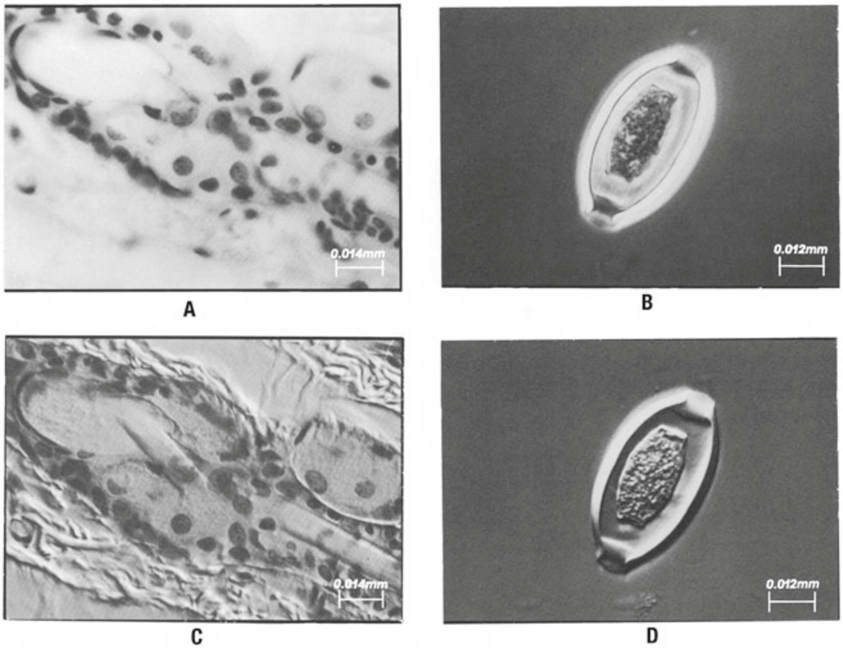


Fig. 30.5 Examples of the contrast obtained with (**a**) bright field, (**b** and **c**) phase contrast, and (**d**) Hoffman modulation optical trains (Photographs courtesy of Dr. Robert Hoffman, Modulation Optics, Inc.)

30.3.3 Polarization Microscopy

A polarized light source can be produced by placing a polarizer in the condensing system of a microscope. Certain objects possess the property of *form birefringence* and will only pass plane-polarized light when the light is parallel to the long axis of the particles. When polarized light is passed through materials that possess form birefringence at an angle of 45°, the light will be resolved into parallel and perpendicular components. A polarization microscope contains two polarizers, one in the condenser and a second between the objective and the eyepiece, called *the analyzer* (Fig. 30.6). If these two polarizers are turned at 90° angles to each other, no light passes and a dark field is observed. If an object possessing form birefringence is placed on the stage of a such a microscope, it will be invisible so long as its long axis is oriented to either the condensing or the analyzer polarizer. However, if the object is placed at a 45° angle to both of the polarizers, it will pass some light and will appear bright against the dark field. An object composed of long molecules in a parallel array or of stacked discs would appear bright at 45° because either of these internal structures will be birefringent. The parallel structure will be positively birefringent and the stacked structure will be negatively birefringent. This

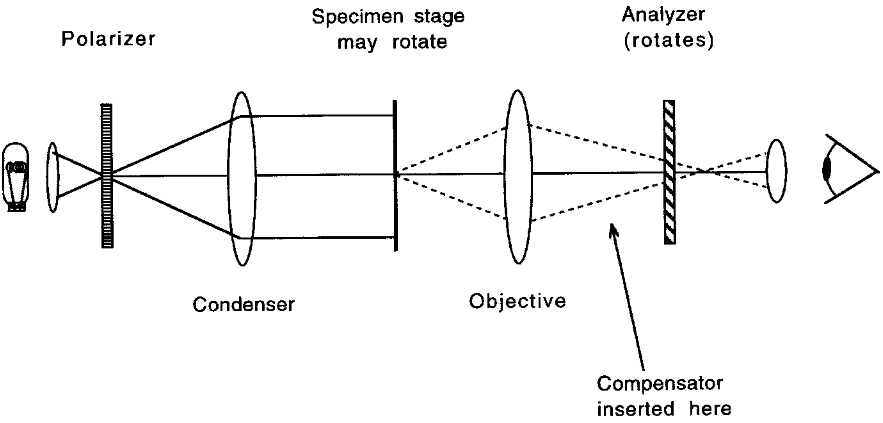


Fig. 30.6 Schematic of the optical train of a polarization microscope

internal structure can be inferred if the sign of the birefringence can be determined. Birefringence occurs when the refractive index of parallel polarized light is different from that of perpendicular polarized light. In the case of positive birefringence the parallel light has a lower index of refraction, and so light travels faster in this direction than in the perpendicular orientation. If a material of known birefringence and direction such as mica is interposed, the velocities can be equalized and the object will have minimum brightness. In the process, use of such a *compensator* will provide the information necessary to indicate the positive or negative birefringence of the material under investigation. Some of the uses of polarizing microscopy in biological investigations are summarized in Table 30.2.

Table 30.2 Uses of polarizing microscopy in biological investigations

Investigation	Example
Orientation of molecules by birefringence	Imaging of stacked molecules in muscle cells, chloroplasts, and retinal rod cells; DNA fiber orientation; myelin protein–lipid structure
Visualization of helical arrays	Amyloid fibril identification (with Congo red stain); inclusion bodies in tobacco mosaic virus
Measuring dichroism	Orientation of heme in hemoglobin; orientation of purine–pyrimidine arrays in chromosomes

30.3.4 Histochemistry

Because biological systems are not only multiphasic but are non-homogeneous systems, it is important to know the location of chemical components within the system. One of the earliest applications of microscopy in biological systems was to combine visualization with the tendency of certain parts of a tissue or a cell to bind a colored dye or undergo a color reaction. A particular chemical species could then be then linked to a specific region of the cell or tissue. This application of chemical analysis to microscopic samples is called *histochemistry*. Histochemical techniques serve two purposes: (1) the color reactions allow the cell to be seen easily under the microscope, and (2) knowledge of the color chemistry and cell biology gives information about the location of a chemical species and the biochemical functioning of the cell. In general one of the limitations of histochemical techniques is that the cells must be killed and their cellular components anchored and immobilized in space (a process called *embedding* or *fixing*) prior to staining. In tissues in which there is a distinct geometry within or among the cells that must be preserved, the tissues are infiltrated with paraffin wax or plastics before being cut thinly into sections, and mounted on microscope slides for staining. Alternatively, the geometry of a single cell can often be preserved by simply heat fixing the cells to a glass slide. A limited number of vital stains are known and are used in living systems, usually to determine cell viability. For example, trypan blue is often used as a vital stain because a living intact cell will exclude the dye from its cytosol. When a cell dies, the trypan blue can breach the cell membrane thus labeling the cells. This dye therefore stains the dead cells in a sample, and if the sample is concurrently examined under phase and bright field conditions a ratio of living or dead cells can readily be calculated. Table 30.3 summarizes some of the histochemical stains commonly used in light microscopy.

Table 30.3 Histochemical stains

Stain	Reaction	Examples
Basic dyes	React with acidic components	
Hematoxylin		Nucleic acids stain blue
Methylene blue		Acid mucopolysaccharides
Acid dyes	React with basic components	
Eosin (acid dye)		Collagen
Orange G		Basic components of proteins in cytoplasm
Acid fuchsin		
Metal impregnation	Impregnates tissues	
Silver		Neurons stain black
Gold		Collagen stains dark brown
		Reticular fibers stain black

30.3.5 Fluorescence Microscopy

The use of a stain that fluoresces can be combined with the ability of the microscope to visualize geometrical arrangements to provide a highly sensitive and multi-modal tool. The optical problem in the fluorescent microscope is to provide an intense monochromatic excitatory light source such that the light emitted (always of a longer wavelength) can be seen. A variety of complicated schemes have been used but today the epi-fluorescent light train is most commonly used (Fig. 30.7). A related technique is the confocal fluorescent microscope. An ever widening array of dyes and fluorescent probes is available that can be used to quantitate components, measure intracellular pH, ions, and redox potential as well as respond to the transmembrane potentials in cells. All of these techniques allow for real-time observation of cellular function in terms of potential energy and force gradients and kinetics. A selection of these techniques is summarized in Table 30.4.

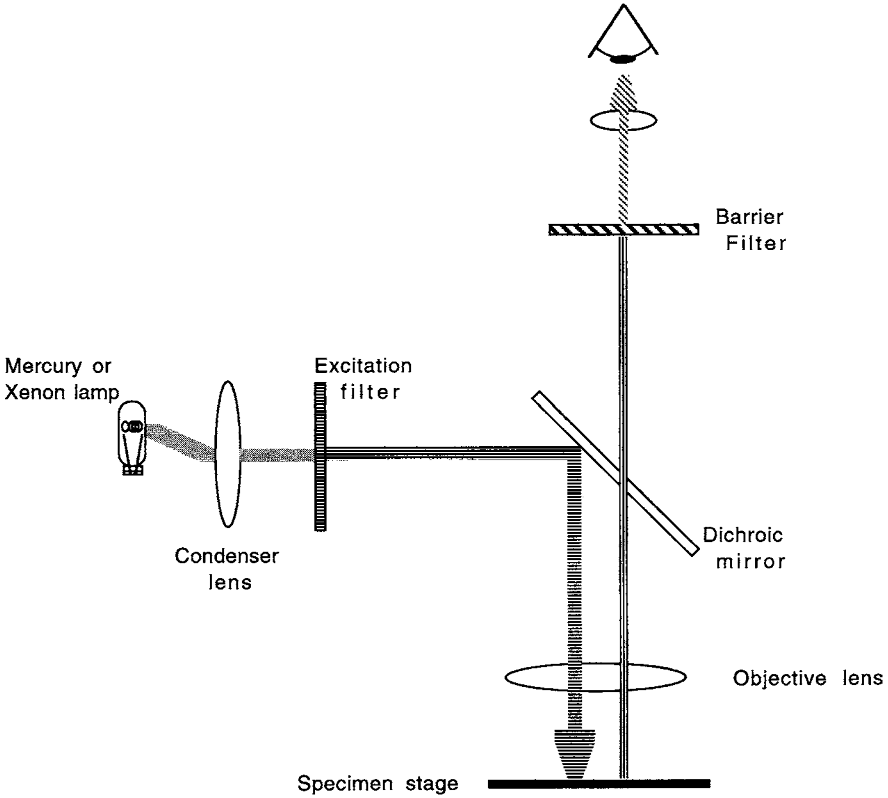


Fig. 30.7 Schematic of the optical train of an epifluorescent microscope

Table 30.4 Fluorescent probes used in cellular studies

Fluorochrome	Excitation (nm)	Emission (nm)	Application
Fluorescein isothiocyanate (FITC)	495	520	Conjugated to immuno-globulins and used as an immuno stain
Lucifer yellow	428	534	Diffuses through gap junctions.
Diethyloxacarbocyanine iodide (DiO)	482	511	Lipophilic, accumulate in membranes with a voltage dependent partioning
Diethyl-3,3,3′,3′-tetra-methylindocarbocyanine iodide (DiI)	549	568	
NAD(P)H	364	4−500	Autofluoresces in living cells. Can be used to indicate cellular redox state
Fura-2	340/380	510−550	Ratiometric Ca^{+2} indicator
Indo-1	365	405/495	Ratiometric free Ca^{+2} indicator
BCECF	495/440	520/550	Ratiometric pH indicator
Di-4-ANEPPS	440/505	640	Ratiometric voltage sensitive indicator

30.4 Scanning Probe Microscopy Creates an Image of a Structures by Interactions on a Molecular Scale

Up to this point, all of the visualization techniques that we have been discussing generate images secondary to the interaction of light with the structural elements of a biological state space. A relatively new set of techniques provides images of a state space without depending on the diffraction of light in the state space. Scanning probe microscopy refers to a growing array of techniques that employ a very small dimension probe that is mechanically scanned back and forth across a surface. The probe can be made sensitive to a variety of forces and can create an image of the force interaction between the probe tip and the surface of the state space under investigation. This technique also allows the optical visualization of the state space but at resolutions not limited by the Abbe diffraction limit given in Eq. (30.1). The concepts underlying scanning probe microscopy are quite straightforward. All forms of scanning probe microscopy use an atomically small probe tip that interacts with the specimen. The tip explores a three-dimensional state space using the interaction force between the tip and the specimen as the observable for that state space. The tip is scanned in an *x*–*y* raster pattern and at each coordinate point in the *x*–*y* plane the interactional force versus the *z*-dimension coordinate is mapped. The resulting *x*, *y*, *z* map gives a force picture that varies in the *z*-dimension. The *z*-dimension force is determined by a feedback circuit that strives to maintain a reference force by moving the probe back and forth in the z dimension at each *x–y* point. These data are then processed by a computer and a pseudo image of the *x, y, z* space is created by a visualization program. A better sense of the application and dimensions of the analysis is gained by considering several of the various scanning probe modalities.

30.4.1 Scanning Tunneling Microscopy

Binnig and Rohrer introduced the scanning tunneling microscope in 1982 and won the Nobel prize in 1986 for this invention (Fig. 30.8). Though it has limited direct applicability in biological research at this time, it is the prototype of scanning probe microscopy and is likely to play a significant role in the future in biological questions, especially in the questions of charge transfer and bioelectrochemistry.

The scanning tunneling microscope takes advantage of the phenomena of quantum mechanical tunneling that we discussed in Chapter 8. Because of the tunneling phenomenon, the electronic distribution of atoms in both the sample and the probe tip spread into space around the equilibrium position of each atom. The exponential decrease in the charge distribution means that if the probe tip is placed within several angstroms of the surface of the specimen under investigation the electron clouds will overlap. By placing a voltage across the sample and probe, the electrons will tunnel

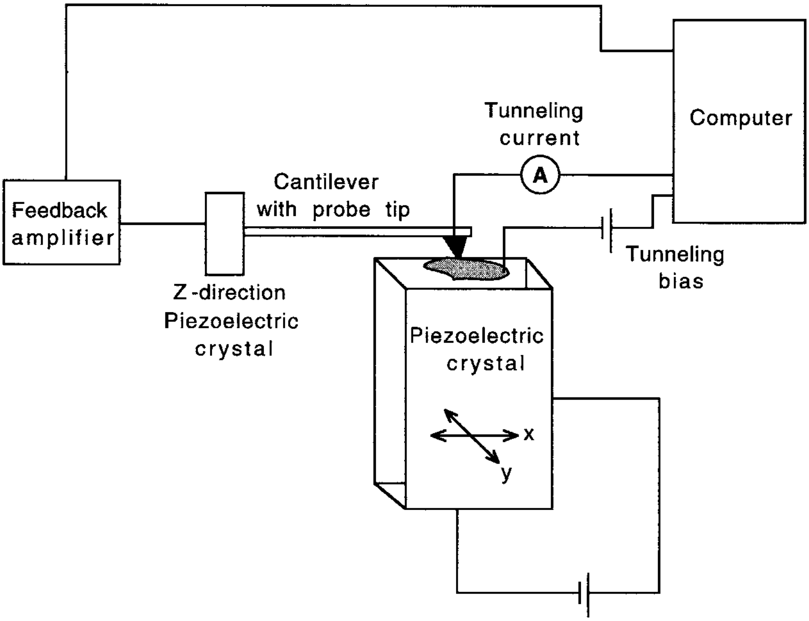


Fig. 30.8 Schematic of the arrangement of the scanning tunneling microscope

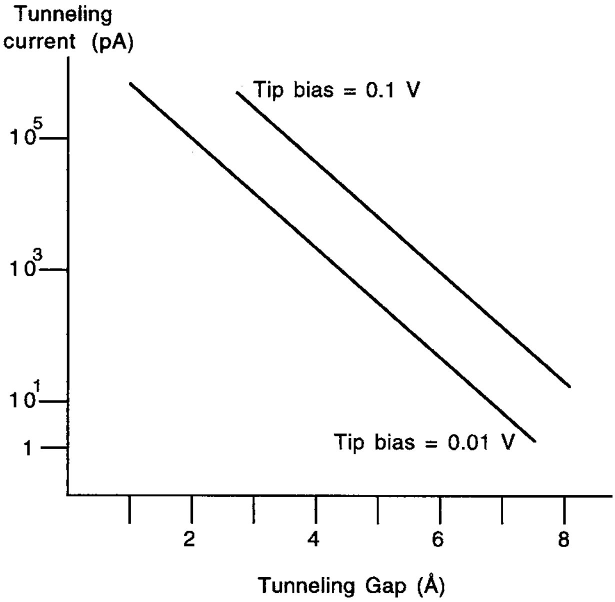


Fig. 30.9 Tunneling current versus distance in the scanning tunneling microscope

across the apparent gap and a small (on the order of picoamperes) but distinct current can be measured. Since the potential field in which the electrons will tunnel is held constant, the tunneling current depends on the conductance between the tip and sample which in turn is related to the apparent charge density of the electron distribution. Since the charge distribution varies exponentially with distance, the tunneling current is exquisitely sensitive to any variation in distance between the sample and the probe (Fig. 30.9). A variation in the *z*-axis of even 1 Å can change the measured tunneling current by an order of magnitude. The three-dimensional map generated by the probe can thus effectively represent a high-resolution topographic map of the surface of the sample to within several angstroms, thus measuring on scale of atomic dimension. Practically, when a scanning tunneling image of a surface is performed, the control circuitry attempts to position the probe in the *z*-dimension so that the current remains constant. The distance the probe travels in the *z*-dimension is thus directly related to the topography of the surface at that point. In general STM requires a conductive sample and thus its application to biological questions has been limited to samples that can be made conductive. (For example, by sputter coating with carbon or platinum metal as is commonly done in shadow-casting for electron microscopy.)

While topographic information is easily appreciated in STM the most exciting observables that can be drawn from the technique are a direct consequence of the quantum mechanical description of the electronic charge distribution in atoms. The charge distribution depends on the *local density of states* which describes which of the possible electronic energy levels of an atom or molecule are occupied or unoccupied. Since fundamentally an electron tunneling between the specimen and the STM tip must enter or exit an electronic energy level in the atom being scanned, the tunneling current reflects the occupancy of those energy levels. By varying the potential between the tip and the specimen, the energy of the levels can be related to the local environment. With this type of data the technique can provide fundamental data on the chemical composition, the electronic and magnetic states, the crystalline structure, and the related chemical structure–function interactions at a quantum level. Used in such a fashion, the STM can perform electrochemical analysis and can perform quantum mechanical voltammetry. Voltammetry, as we have seen earlier, is a form of spectroscopic analysis of the electrochemical potential.

30.4.2 Scanning Force Microscopy

The prototypical scanning force microscope, the *atomic force microscope (AFM)*, has a very fine probe tip mounted on a cantilever arm which moves in the *z* direction as the sample is raster scanned in the *x*–*y* dimension. In the AFM, the probe tip is brought within very close proximity to the sample. At relatively long distances, the tip experiences the attractive van der Waals forces, and the cantilever is bent toward the sample. As the tip is moved more closely to the surface a repulsive force secondary to the Pauli exclusion of the outer orbit electrons becomes dominant, and the probe and cantilever are bent away from the surface. These deformations of the cantilever are registered in the *z*-axis feedback circuits by a multi-sectored photocell that senses the reflections of a laser focused on the cantilever head (Fig. 30.10). The force

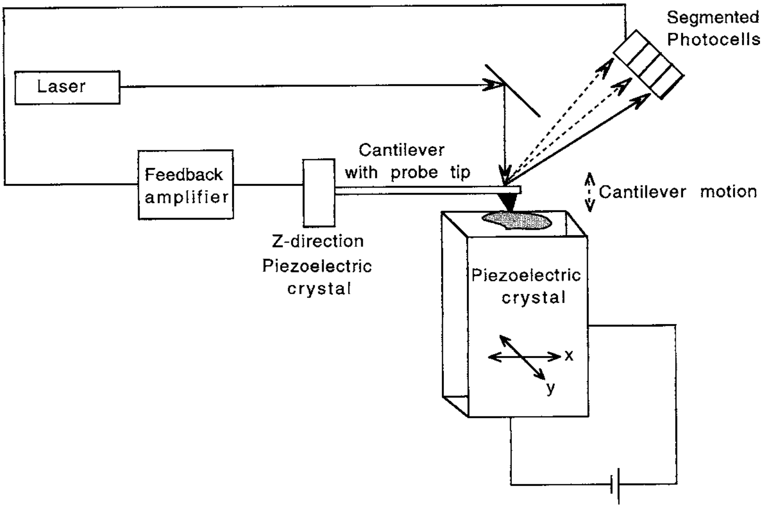


Fig. 30.10 Schematic of the atomic force microscope

acting on the probe can be determined by recognizing that the cantilever is simply a form of spring whose resistance to deformation (restoring force) is the same as the spring constant from Hooke's law. The stiffer the cantilever (higher spring constant) the smaller the displacement for a given attractive or repulsive force and vice versa. Cantilevers also oscillate like springs, and the resonant frequency depends on the spring constant. If a cantilever is put into harmonic motion and the vibrating probe is acted on by another force such as the attractive van der Waals force, the oscillatory motion will be affected, and the modulation of the harmonic motion can provide information about the interaction energy. Both of these mechanical interactions in an atomically small dimension can be used to provide topographical (*z*-axis) and lateral force information (which can reflect hardness and stickiness). While most AFM applications in biological systems have been focused predominantly on generating an image of a macromolecule or biological system like a cell, the techniques have much broader promise. For example by having a ligand attached to the AFM probe, allowing binding of the ligand to its receptor and then pulling the complex apart, the AFM has allowed the force of the receptor–ligand interaction to be directly measured. Furthermore, because the AFM can be used in liquid phase and in samples that require virtually no preparation, it can be used to make measurements in living systems or to observe macromolecular assembly in real time (Fig. 30.11). It is not difficult to appreciate that the cantilever can respond to any force associated with the sample. Whether the probe responds to the force depends on how the probe is constructed. Among the forces that scanning force microscopes have been built to sense are magnetic and electrical fields.

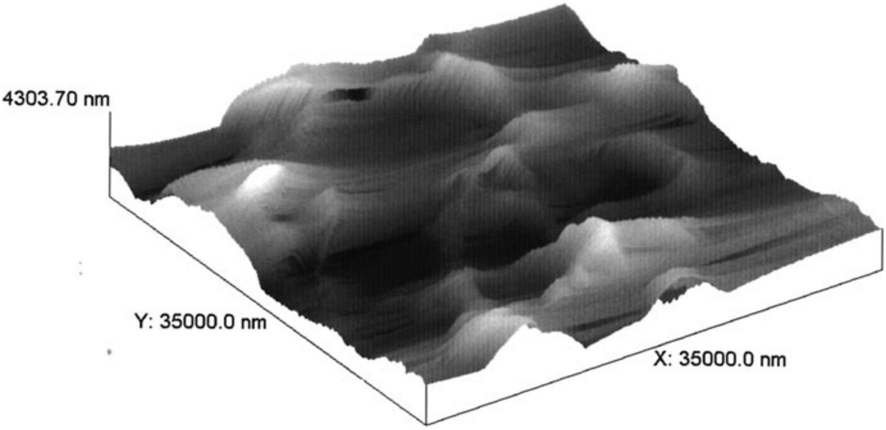


Fig. 30.11 AFM image of an amyloid fibril deposit isolated from a patient with familial amyloidotic polyneuropathy. All amyloid fibrils, regardless of the biochemical or clinical type, share a typical composition at least when studied at the level of the electron microscope. Amyloid deposits are comprised of an array of linear, rigid, non-branching aggregated fibrils that are 7.5–10 nm in diameter but of indefinite length. The fibrils are formed from two or more filaments, each with a diameter of 0.25–0.35 nm that forms a helically twisted structure with a hollow central core. The helical structure gives the amyloid fibril a beaded appearance (Courtesy of Symmetry Research)

30.4.3 Near-Field Optical Microscopy, Outside the Classical Limits

We have demonstrated that there are limits to image resolution obtained by a diffraction technique when using light. The diffraction barrier is considered to be on the order of 1/2 a wavelength of the illuminating light. For a microscope forming an image using green (500 nm) light the best resolution possible would be on the order of 0.2–0.25 μm. Practically, the best resolution is usually closer to 0.5 μm. In 1928, E.H. Synge proposed that the far-field diffraction limit could be overcome, if a light source emanated from an aperture that was much smaller than the wavelength of the illuminating light, and if the light source was brought much closer to the object than the wavelength of the light. The light as such is not focused or diffracted but acts like a lantern illuminating a spot just several feet in front of it. Such a system was not technically feasible in 1928; but in 1972 the theoretical construct was shown to be correct with microwave light. With the advent of the micropositioning technologies that made scanning tunneling microscopy a reality, the near-field optical instrument is now available.

Further Reading

General

Freeman W.J. (1991) The physiology of perception, *Sci. Am.*, **264, 2**:78–85. (Thoughts on the nature of the power of visualization.)

Freifelder D.M. (1982) *Physical Biochemistry: Applications to Biochemistry and Molecular Biology*, 2nd edition. W.H. Freeman, New York. (Nice discussion of applications.)

Newer Light Microscopic Techniques

Betzig E. and Trautman J.K. (1992) Near field optics: microscopy, spectroscopy and surface modification beyond the diffraction limit, *Science*, **257**:189–195.

Howells M.R., Kirz J., and Sayre W. (1991) X-ray microscopes, *Sci. Am.*, **264, 2**:88–94.

Lichtman J.W. (1994) Confocal microscopy, *Sci. Am.*, **271, 2**:40–45.

Scanning Force Microscopy:

General

Hansma C.E., Elings V.B., Marti O., and Bracker C.E. (1988) Scanning tunneling microscopy and atomic force microscopy: application to biology and technology, *Science*, **242**: 209–216.

Nölting B. (2006) *Methods in Modern Biophysics*, 2nd edition. Springer-Verlag, Berlin.

Sarid D. (1994) *Scanning Force Microscopy with Applications to Electric, Magnetic and Atomic Forces*. Oxford University Press, New York.

Wickramasinghe H.K. (1989) Scanned-probe microscopes, *Sci. Am.*, **261, 4**:98–105.

Applications

Bustamante C. and Rivetti C. (1996) Visualizing protein-nucleic acid interactions on a large scale with the scanning force microscope, *Annu. Rev. Biophys. Biomol. Struct.*, **25**:395–429.

Ikonomovic M.D., Armstrong D.M., Yen S., Obcemea C., and Vidic B. (1995) Atomic force microscopy of paired helical filaments isolated from autopsied brains of patients with Alzheimer's disease and immunolabeled against microtubule-associated protein tau, *Am. J. Pathol.*, **147**:516–526.

Rief M., Oesterhelt F., Heymann B., and Gaub H.E. (1997) Single molecule force spectroscopy on polysaccharides by atomic force microscopy. *Science,* **275**:1295–1297.

Yip C.M. and Ward M.D. (1996) Atomic force microscopy of insulin single crystals: direct visualization of molecules and crystal growth, *Biophys. J.*, **71**:1071–1078.

Problem Sets

1. We must recognize that each level of exploration carries a cost versus benefit problem. Perhaps this is best called the "forest or trees" conflict. With greater visualization of detail there is a loss of the overall organization of the system. This is a common problem in microscopy. What are the problems with increasing resolution, histochemical analysis, fluorescent markers. What is gained and what is lost with each of these observational levels. Is one level better than any other?

Chapter 31
Epilogue

We have come to the end of our trail together. A generous amount of time has been given to tying together the background fields of physics, mathematics, and chemistry. The reward for this good effort is that thousands of biological examples of in fields from anatomy to zoology and molecules from α-1-anti-trypsin to zymogen are now accessible in the primary literature. Our trip, you see, has been on a road to a terminal. But this terminal is not the end of the road. It is a port-of-call to new and deeper knowledge. The exploration of systems whether biological, man-made, or purely physical should now has unity. We have left many details to your own discovery on this path we have taken. But now there are very few new trails that you could set upon that will not seem familiar and in a fashion, friendly. Fields that might never have held much interest for you will now beckon. Pick up a copy the *Journal of Physical Chemistry* or browse in *Cell*. There will be work there of interest and you will know how to go about exploring it.

The world of bio-physical science is filled with marvel and surprise. You have earned your ticket. Come. Let us go. It's time to explore.

> I seem to have been only like a boy playing on the seashore, and diverting myself in, now and then, finding a smoother pebble or a prettier shell than ordinary, while the great ocean of truth lay all undiscovered before me.
>
> Isaac Newton

Now Try

Birge R.R. (1995) Protein-based computers, *Sci. Am.*, **272, 3**:90–95. (An interesting use of the interaction between light and bacteriorhodopsin for molecular switching.)

Drexler K.E. (1992) *Nanosystems, Molecular Machinery, Manufacturing and Computation.* Wiley Interscience/Wiley, New York. (The biological curriculum applied to the future of technology.)

Pascher T., Chesick J.P., Winkler J.R., and Gray H.B. (1996) Protein folding triggered by electron transfer, *Science*, **271**:1558–1560. (A paper that explores the protein folding problem using thermodynamics and dynamic electrochemical treatments.)

Regan L. (1993) The design of metal-binding sites in proteins, *Annu. Rev. Biophys. Biomol. Struct.*, **22**:257–281. (The future ahead will employ a knowledge of biophysical chemistry to design and engineer biological systems for biomedical and industrial uses.)

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_31, © Springer Science+Business Media, LLC 2010

Roberts T.J., Marsh R.L., Weyand P.G., and Taylor C.R. (1997) Muscular force in running turkeys: the economy of minimizing work, *Science*, **275**:1113–1115. (An elegant application of biophysics to the mechanics of running. Does a turkey pogo stick or pump its way across the barnyard?)

Scrödinger E. (1992) *What is Life?* Cambridge University Press, Cambridge. (Reprint of the 1944 book by a quantum physicist looking at the biological system. Time has shown that many of Scrödinger's speculations were in error but it is a useful example of cross-disciplinary thinking.)

Chapter 32
Physical Constants

Atomic mass unit	U	1.661×10^{-27} kg
Avogadro's number	N_A	6.022×10^{23} mol^{-1}
Bohr magneton	μ_B	9.27×10^{-24} J/K
Boltzmann constant	K	1.381×10^{-23} J/K
Electron rest mass	M_e	9.100×10^{-31} kg
Elementary charge	E	1.602×10^{-19} C
Faraday constant	F	9.6485×10^{4} C/mol 2.306×10^{4} cal/mol/eV
Neutron mass	M_n	1.673×10^{-27} kg
Permeability of free space	μ_o	$4\pi \times 10^{-7}$ T m/A
Permittivity of vacuum	ε_o	8.854×10^{-12} C^2/N-m^2
Physical constants		
Planck constant	H	6.626×10^{-34} J s
Proton rest mass	M_p	1.673×10^{-27} kg
Speed of light (in vacuum)	C	2.990×10^{8} m/s
Universal gas constant	R	8.314 J/K-mol

Conversions

1 Joule = 1 Newton meter
1 atmosphere = 1.01325×10^5 Pa (Pascal)
1 liter = 1×10^{-3} m^3

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6_32, © Springer Science+Business Media, LLC 2010

Appendix A
Mathematical Methods

Mathematics is the natural language of biophysical chemistry even though many biologists and chemists are uncomfortable with that language. The key to the modeling procedures, whether descriptive, explanatory, or experimental, used in biophysical investigations is expression of the relationships in mathematical terms both quantitative and qualitative. In this appendix, many of the mathematical tools and results used in this book are briefly reviewed.

A.1 Units and Measurement

A primary goal of biophysical investigations is to describe the dimensions and shape of a biochemical system by measurement and quantification. Dimensional analysis and the uncertainties inherent in empirical measurement are important practical problems in biophysical studies. A physical quantity is described as a numerical multiple of a standardized unit. The SI system of measurement is now preferred, though in a wide variety of biologically oriented literature still uses non-SI standard units. An idealized goal of a standard system of measurement and specifically one of the goals of the metric system is the definition of a single unit from which all other quantities can be derived. This goal has not been realized today. The SI system now consists of seven fundamental or base units, all other units are derived from these base units. The fundamental quantifiable classes of the SI system are length, mass, time, electric current, thermodynamic temperature, amount of substance, and luminous intensity. There are internationally agreed upon primary reference standards for the description of each of these units. These references are periodically redefined as methods of increased accuracy and technologies to make primary reproduction of the standard quantity become decentralized. For example, the meter, originally defined in terms of the arc of meridian that passed through Dunkerque, France and Barcelona, Spain allowed a platinum–iridium rod to be constructed that was kept at the French National Archives in France since 1799. This reference was subsequently redefined as 1.65076373×10^6 times the wavelength of krypton-86 emitting at and in 1983 it was again redefined as the length of the path of light traveling in a vacuum in 1/299,792,458 s. Table A.1 summarizes this system of measurement. Table A.2 provides some of the common conversion factors.

P.R. Bergethon, *The Physical Basis of Biochemistry*,
DOI 10.1007/978-1-4419-6324-6, © Springer Science+Business Media, LLC 2010

Table A.1 International System (SI) of units

Measurement class	Base unit	Symbol
Length	meter, m	l
Mass	kilogram, kg	m
Time	second, s	t
Electric current	ampere, A	i
Thermodynamic temperature	kelvin, K	T
Amount of substance	mole, mol	n
Luminous intensity	candela, cd	I_v
Derived class	**Derived name**	**Derived unit**
Force	newton, N	kg m s^{-2}
Pressure	pascal, Pa	kg m^{-1} s^{-2}
Energy, work, heat	joule, J	kg m^2 s^{-2}
Power, radiant flux	watt, W	kg m^2 s^{-3}
Electric charge	coulomb, C	A s
Electric potential	volt, V	kg m^2 s^{-3} a^{-1} (J A^{-1} s^{-1} or J C^{-1})
Electrical resistance	ohm, Ω	kg m^2 s^{-3} A^{-2} (V A^{-1})
Electrical capacitance	farad, F	A s V^{-1} or (m^{-2} kg^{-1} s^4 A^2)
Frequency	hertz, Hz	s^{-1} (cycles per second)
Magnetic flux density	Tesla, T	kg s^{-2} A^{-1} (N A^{-1} m^{-1})

Table A.2 Conversion factors

Quantity	Unit	SI equivalent
Length	angstrom	1.000×10^{-10} meter
Force	dyne	1.000×10^{-5} newton
Pressure	atmosphere	1.013×10^{5} pascal
	bar	1.000×10^{5} pascal
	torr (mmHg at 273 K)	1.333×10^{2} pascal
Energy	electronvolt	1.602×10^{-19} joule
	wave number (m^{-1})	1.986×10^{-25} joule
	calorie (thermochemical)	4.184×10^{0} joule
	erg	1.000×10^{7} joule
Electric charge	e.s.u. (electrostatic unit of charge)	3.336×10^{-10} ampere
Magnetic flux density	gauss	1.000×10^{-4} tesla

A.2 Exponents and Logarithms

The system of numbering based on ten is the original digital numbering systems because of the nearly universal possession by humans of ten digits on their hands. In many cases, another numbering system might be more convenient. For example the binary system based on the base 2 system is preferred when dealing with digital computers and in Boolean logic. An interesting case could be made for the use of a base 4 system when dealing with genomic codes.

Numbers can be arranged in a particular order in which the subsequent number depends on the application of a specific mathematical function to the preceding number. If the operating function increases or decreases each number in the set by a fixed amount, an arithmetic series is the result.

$$\text{(a)}\ \ 1, 2, 3, 4, 5, 6, \ldots, (x + 1)$$

A geometric progression is generated if each preceding term is multiplied by a fixed number to get the next term.

$$\text{(b)}\ \ 1, 2, 4, 8, 16, 32, 64, \ldots, (2x)$$

$$\text{(c)}\ \ 1, 5/6, 25/36, 125/216, 625/1296, \ldots, 5/6x$$

$$\text{(d)}\ \ m, mr, mr^2, mr^3, mr^4, \ldots, mr^{x+1}$$

The properties of series are extremely useful in solving a variety of problems. For now, inspection of the series (b) and (d) above provides a good reminder of the valuable properties of exponents and logarithms. If the constant r in (d) is chosen as the multiplication factor in the previous series (b), i.e., (d) is expressed in base 2 and these series are arranged side by side, it is evident that addition of exponents in series (d) is equivalent to multiplication of terms in series (b).

(a)	(b)	(d)
1	2	2^0
2	4	2^2
3	8	2^3
4	16	2^4
5	32	2^5
6	64	2^6
7	128	2^7
8	256	2^8

For example, multiplying 8 times 32 in (b) is the same as adding the exponents 3 and 5 in (d). This use of logarithms in base 2 can be written as

$$2^n = N$$

and

$$\log_2 N = n$$

which defines the relationship of n as the logarithm of N in base 2. This is an example of the more general case for logarithms which can be written as

$$\log_b N = n, \text{ and, } b^n = N$$

As the example above showed, logarithms are simply exponents written without the base and hence the rules of exponents can be applied to them. In algebra, b^n was defined as b taken n times and therefore it can be shown that

$$b^n + b^m = b^{n+m} \text{ and}$$

$$(b^n)^m = b^{nm}$$

Exponents of value zero, negative, and fractional value are written as

$$b^0 = 1$$

$$b^{-n} = 1/b^n$$

$$b^{n/m} = (m\sqrt{b})^n$$

These rules governing exponents can now be used to define rules of logarithms. Some of the most important include the following:

$$\log_b NM = \log_b N + \log_b M$$

$$\log_b(N/M) = \log_b N - \log_b M$$

$$\log_b M^k = k \log_b M$$

The relationship between logarithms from two different bases can be found. Let N be the logarithm for the following systems:

$$N = a^x \text{ and } N = b^y$$

The following equality is written:

$$a^x = b^y$$

which can be expressed in logarithmic form in base a

$$\log_a a^x = \log_a b^y$$

x and y are by definition equal to $\log_a N$ and $\log_b N$, respectively, substituting gives this result

$$\log_a N \ \log_a a = \log_b N \ \log_a b$$

Since $\log_a a$ must be equal to unity, this equation simplifies to

$$\log_a N = \log_b N \ \log_a b$$

This equation transforms the logarithm of one base (a) into the logarithm of another base (b).

In most biophysical chemical applications, the most common logarithmic systems used are those in base 10 and the natural base system e.

A.3 Trigonometric Functions

The trigonometric functions find wide application in the biophysical sciences. Trigonometry is valuable because it provides us with a method of working with functions that periodically repeat. Such functions map out onto state space in a circular fashion. We start by considering the equation of a circle such that

$$x^2 + y^2 = 1$$

The set of points that will satisfy this equation will map out a circle in Cartesian space with a unit radius of 1 and with a center point at the origin of the Cartesian system. We are interested in ways in which we can enumerate any element of the solution set by locating a point on the circle which represents the x, y pair. One method of enumeration is to measure the length of the arc between the point of interest and the $+x$ axis position of 1. Because this is a circle with a unit radius of 1 and the circumference of any circle is given by $2\pi r$, this circle's circumference is simply 2π. We can identify any element (x, y) by a measurement between 0 and 2π which represents the length of the arc. Alternatively we could have indicated a point on the circle with a measurement of the angle made between the $+x$ axis and the unit radius as it sweeps through the space. We can equate the value of this angle with the measure of the distance of the arc it sweeps. Since the arc must be between 0 and 2π in length, the angle also varies between 0 and 2π. When we employ this mapping approach the angle is given in *radians* and 2π radians equals a complete circle. Initially, the use of radians is always a somewhat cumbersome for the student because we tend to use angle measurement in degrees on a routine basis throughout our non-scientific lives. Radian measurements are a more natural unit and using them is particularly useful and simplifying especially when taking the derivatives of trigonometric functions. We can easily convert between radians and degree measurements for angles be equating the two facts that the unit vector sweeps out an arc of 2π radians or 360° to make a complete circle. Therefore

$$2\pi \text{ radians} = 360^\circ$$

$$1 \text{ radian} = \frac{180}{\pi} \text{ degrees}$$

The six basic trigonometric functions are derived from the unit circle and are as follows:

$$\sin\theta = \frac{y}{r} = \frac{y}{\sqrt{x^2+y^2}}$$

$$\cos\theta = \frac{x}{r} = \frac{x}{\sqrt{x^2+y^2}}$$

$$\tan\theta = \frac{\sin\ \theta}{\cos\ \theta} = \frac{y}{x}$$

$$\cot\theta = \frac{\cos\ \theta}{\sin\ \theta} = \frac{x}{y}$$

$$\sec\theta = \frac{1}{\cos\ \theta} = \frac{r}{x} = \frac{\sqrt{x^2+y^2}}{x}$$

$$\operatorname{cosec}\theta = \frac{1}{\sin\ \theta} = \frac{r}{y} = \frac{\sqrt{x^2+y^2}}{y}$$

Graphs of these functions are illustrated in Fig. A.1.
There are a number of important trigonometric identities that find usage in biophysical studies. Among these are

$$\begin{aligned}
&\sin^2\theta + \cos^2\theta = 1\\
&\sin(-\theta) = -\sin\theta\\
&\cos(-\theta) = \cos\theta\\
&\tan(-\theta) = -\tan\theta\\
&\sin(A+B) = \sin A\cos B + \cos A\sin B\\
&\sin(A-B) = \sin A\cos B - \cos A\sin B\\
&\cos(A+B) = \cos A\cos B - \sin A\sin B\\
&\cos(A-B) = \cos A\cos B + \sin A\sin B\\
&\tan(A+B) = \tfrac{\tan A + \tan B}{1 - \tan A \tan B}\\
&\tan(A-B) = \tfrac{\tan A - \tan B}{1 + \tan A \tan B}\\
&\sin 2\theta = 2\sin\theta\cos\theta\\
&\cos 2\theta = \cos^2\theta - \sin^2\theta
\end{aligned}$$

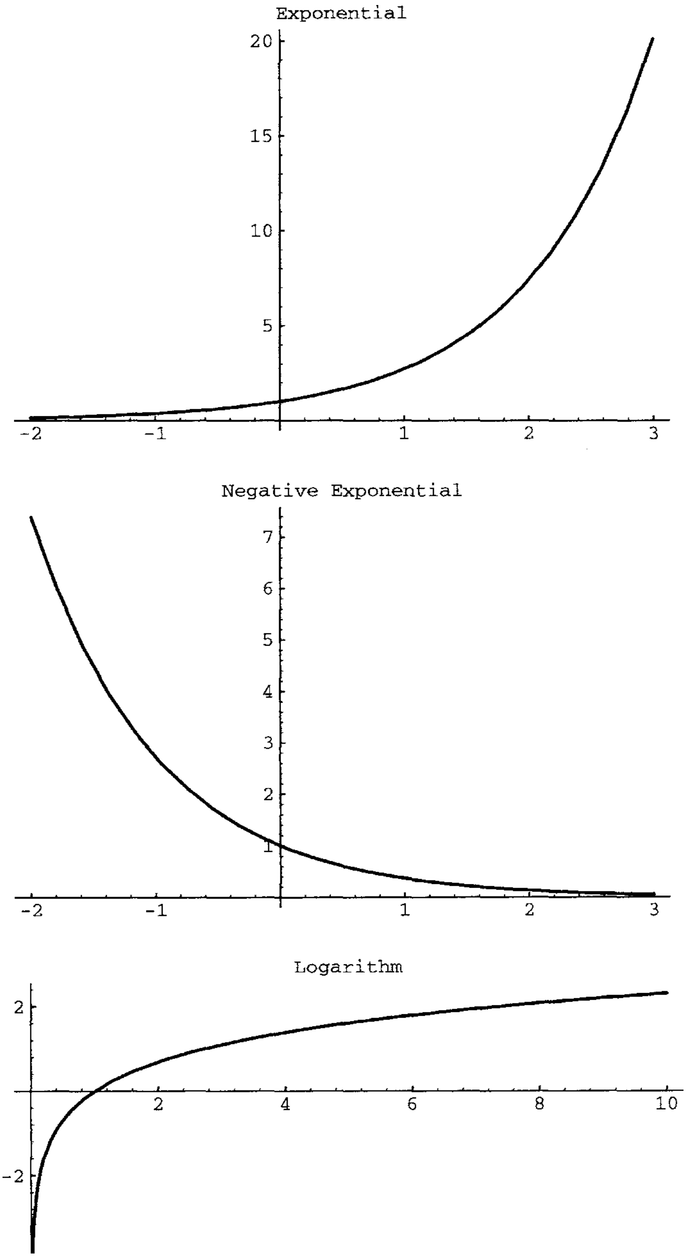


Fig. A.1 Graphs of the basic transcendential functions: **A,** *Exponents*: exponential, $y = e^x$; negative exponential, $y = e^{-x}$; logarithmic, $y = \ln x$. **B**, *Trigonometric*: sine, $y = \sin x$; cosine, $y = \cos x$; tangent, $y - \tan x$. **C**, *Hyperbolic*: sinh, $y = \sinh x$, $\sinh x = \frac{1}{2}\left(e^x - e^{-x}\right)$; cosh, $y = \cosh x$, $\cosh x = \frac{1}{2}\left(e^x - e^{-x}\right)$; tanh, $y = \tanh x$, $\tanh x = \frac{\sinh x}{\cosh x}$; cosecnth, $y = \mathrm{csch} x$, $\mathrm{csch} x = \frac{1}{\sinh x}$; secanth, $y = \mathrm{sech} x$, $\mathrm{sech} x = \frac{1}{\cosh x}$; cotangenth, $y = \cot x$, $\coth x = \frac{\cosh x}{\sinh x}$

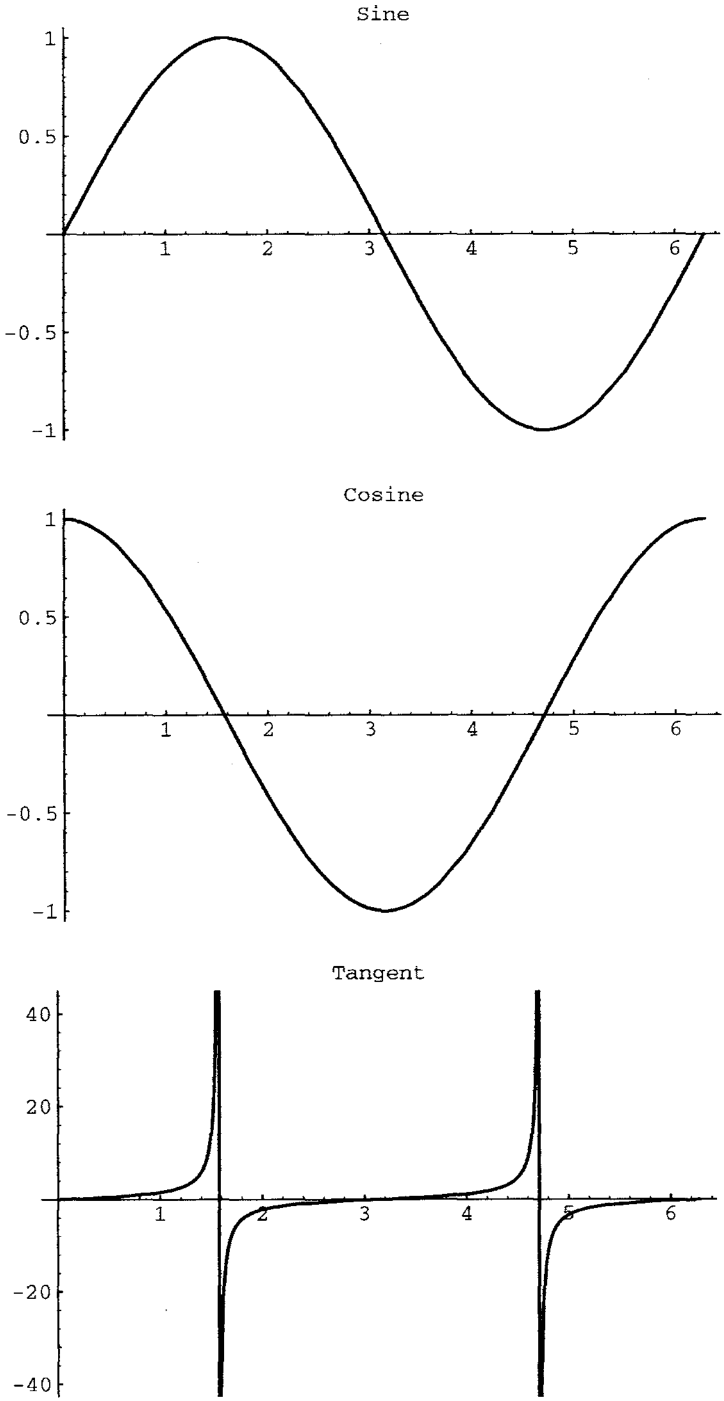


Fig. A.1 (continued)

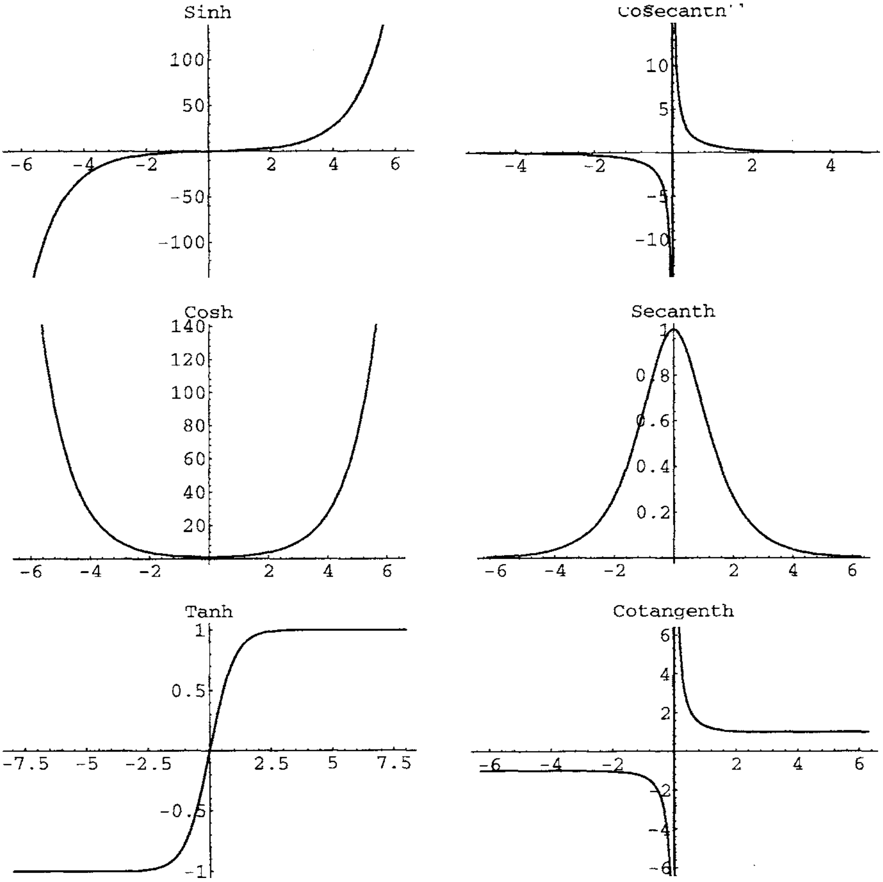


Fig. A.1 (continued)

A.4 Expansion Series

A power series results from the functions whose terms are not constants but rather those terms are functions of x such that the nth term is a constant times x^n or $(x{-}a)^n$ where a is a constant.

An appropriate series can be used to evaluate a variety of expressions. We will use

$$e^x\,(\text{for}\quad\text{all}\,x) = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \frac{x^4}{4!} + \cdots$$

$$e^{-x}\,(\text{for}\;\text{all}\,x) = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \cdots$$

and the *geometric series*:

$$1+x+x^2+x^3+x^4+\cdots=\frac{1}{1-x}\text{for}-1<x<1$$

A.5 Differential and Integral Calculus

The techniques of calculus are widely used in biophysical problems usually as a method of simplifying and abstraction in a complicated model. It is surprising to many students and practitioners of biological sciences that mathematics can be promoted as a simplifying solution but with practice and familiarity it is a powerful shorthand. In fact many of the basic problems are seen over and over so in fact a level of comfort with just several basic differentials (and integrals) can be a very adequate and powerful tool. Table A.3 lists most of the important differentials and integrals used in this text.

Table A.3 Table of common derivatives and integrals used in this book

A.	$\int ax^n dx = \frac{ax^{n=1}}{n+1}$	$(n\neq 1)$
B.	$\int ax^{-1} dx = a\int \frac{dx}{x} = a\ln x$	
C.	$\int_{x_1}^{x_2} ax^n dx = \frac{a\left(x_2^{n+1}-x_1^{n+1}\right)}{n+1}$	$(\mathrm{n}\neq 1)$
D.	$\int_{x_1}^{x_2} ax^{-1} dx = a\left(\ln x_2 - \ln x_1\right) = a\ln\frac{x_2}{x_1}$	
E.	$\frac{d\sin a\theta}{d\theta} = a\cos a\theta$	
F.	$\frac{d\cos a\theta}{d\theta} = -a\sin a\theta$	
G.	$\frac{d(e^{ax})}{dx} = ae^{ax}$	
H.	$\frac{d\left(e^{ax^2}\right)}{dx} = 2axe^{ax^2}$	

A.6 Partial Differentiation

Often we will be interested in a system in which many variables are linked and will change as a function of one another. When evaluating such a system the principles of partial differentiation are employed in which one observable is allowed to vary while all others are held constant. The most practical way of working with partial differentials is to recognize their physical meaning as a path along the intersection between several planes drawn by graphing a mathematical function. The overall system is conceived as a multi-dimensional set of planes and a single element of the partial differential equation is a function of the two variables defining that plane. The techniques of partial differentiation represent the set of rules whose purpose is to relate the linkages between the variables on one plane to variables on another plane (and other functions) in the overall system. Partial differential equations are often clothed in great mystery but in fact if viewed as a decoding ring for a treasure map, the mystery can also become an enjoyable puzzle.

The rules for partial differential decoding are

1. *f* is a function of *x* and *y*. When *x* and *y* are changed by *dx* and *dy*, *f* must also change by *df*. We can write this in equation form:

   $$df = \left(\frac{\partial f}{\partial x}\right)_y dx + \left(\frac{\partial f}{\partial y}\right)_x dy$$

   Derivatives can be taken in any order:

   $$\left[\frac{\partial(\frac{\partial f}{\partial y})_x}{\partial x}\right]_y = \left[\frac{\partial(\frac{\partial f}{\partial x})_y}{\partial y}\right]_x$$

2. In many cases *x* and *y* will depend on another variable *z*. We may need to find out how *f* changes when *x* is changed assuming that *z* is held constant, so

   $$\left(\frac{\partial f}{\partial x}\right)_z = \left(\frac{\partial f}{\partial x}\right)_y + \left(\frac{\partial f}{\partial y}\right)_x \left(\frac{\partial f}{\partial x}\right)_z$$

3. If *z* is a function of *x* and *y*, as is the case in a linkage such as an equation of state, then the inverse of a function can be found as follows:

   $$\left(\frac{\partial x}{\partial y}\right)_z = \frac{1}{\left(\frac{\partial y}{\partial x}\right)_z}$$

4. We are often interested in the relationships between the elements in equation of state. When *x*, *y*, and *z* are related we can find the permutations by

   $$\left(\frac{\partial x}{\partial y}\right)_z = -\left(\frac{\partial x}{\partial z}\right)_y \left(\frac{\partial z}{\partial y}\right)_x$$

5. Euler’s chain rule results from a combination of rule 3 and rule 4:

   $$\left(\frac{\partial x}{\partial y}\right)_z \left(\frac{\partial z}{\partial x}\right)_y \left(\frac{\partial y}{\partial z}\right)_x = -1$$

A.7 Vectors

Usually when we use numbers and algebraic expression we are concerned with the magnitude or amount that the number represents. Such expressions are scalar quantities. Many physical quantities have both magnitude and direction, and these are treated by vectors. Force, acceleration, and electric field are all examples of vectors. Vectors are often represented in *bold* type, a convention we will follow. When discussing direction a coordinate system is required and we have already discussed a variety of common and convenient coordinate systems. It is worth noting that complex numbers, which have components that can be described in terms of two orthogonal planes, the real and imaginary planes, are vectors and vector techniques apply to these expressions.

Generally we conceive of vectors as being an arrow drawn in a coordinate space. The length between its origin and the point of the arrow (point A) is its magnitude and its direction is from the origin to point A. A vector with a magnitude of a single unit is called a unit vector. Unit vectors are popular because the algebra is simplified and they are self-normalizing. If a unit vector can be chosen whose origin is the actual origin point of a coordinate system, the math is again simplified because the coordinates of the origin are all 0 and the magnitude of the vector is 1. Such a vector is the radius vector, **r**. We usually write vectors in terms of their components. A vector can be described in terms of three mutually perpendicular axes X, Y, and Z, each of which is characterized by a unit vector, respectively, **i**, **j**, and **k**. Thus

$$\mathbf{r} = x\mathbf{i} + y\mathbf{j} + z\mathbf{k}$$

There are specific rules that allow combination of vectors. We will summarize them here. These rules can be performed graphically or analytically. The experienced scientist will often sketch out a vector problem graphically before attempting an analytical solution and many times when an analytical solution is too difficult to solve precisely.

A.7.1 Addition and Subtraction

Two vectors are added to give a third vector:

$$\mathbf{A} + \mathbf{B} = \mathbf{C}$$

$$\mathbf{A} + (-\mathbf{B}) = \mathbf{D}$$

We draw A in space and then place the tail of **B** at the point of **A**. The sum of the two vectors is the resultant **C**. Subtraction is simply the addition of the negative vector. Both addition and subtraction of vectors is commutative and the magnitude and direction of the vector is independent of the coordinate system. As a consequence, any point in space can be equivalently treated as the origin.
Writing the vectors in terms of their components allows addition and subtraction to be performed analytically.

$$\mathbf{A} = \mathrm{A}_x\mathbf{i} + \mathrm{A}_y\mathbf{j} + \mathrm{A}_z\mathbf{k}$$

$$\mathbf{B} = \mathrm{B}_x\mathbf{i} + \mathrm{B}_y\mathbf{j} + \mathrm{B}_z\mathbf{k}$$

therefore

$$\mathbf{C} = (\mathrm{A}_x + \mathrm{B}_x)\mathbf{i} + (\mathrm{A}_y + \mathrm{B}_y)\mathbf{j} + (\mathrm{A}_z + \mathrm{B}_z)\mathbf{k}$$

A.7.2 Magnitude of a Vector

The magnitude of a vector can be found by measurement of each component graphically or by the trigonometric relationship:

$$r = \sqrt{x^2 + y^2 + z^2}$$

This is generally applicable to any vector so the magnitude of **A** can be written as

$$|\mathrm{A}| = \sqrt{\mathrm{A}_x^2 + \mathrm{A}_y^2 + \mathrm{A}_z^2}$$

A.7.3 Multiplication

Two types of vector multiplication can be performed. Dot or scalar multiplication (**A** • **B**) gives a scalar product of two vectors. The cross or vector product (**A x B**) results in a vector.

The scalar product is defined as the product of the length of *A* times the length of the projection of *B* onto *A*. This product commutes. Analytically we write as

$$A \bullet B = AB\cos\theta$$

where A and B are the magnitudes of **A** and **B** and θ is the angle between **A** and **B**. θ is always written so that it is less than 180°. When two vectors are perpendicular, $\cos\theta = \cos 90 = 0$ which makes $\mathbf{A} \bullet \mathbf{B} = 0$. These vectors are called *orthogonal*. In terms of components:

$$\mathbf{A} \bullet \mathbf{B} = A_x B_x + A_y B_y + A_z B_z$$

The vector product is defined as

$$\mathbf{AxB} = nAB \sin\theta$$

n is a unit vector that is perpendicular to both **A** and **B**. The determination of the direction of **n** is accomplished through the use of the *right-hand rule*. This rule is applied by placing the ulnar surface of the palm (the side opposite the thumb) pointing in the direction of the **A** vector and curling the fingers toward the **B** vector. The thumb points perpendicular to the hand and in the direction of **n**. The vector product does not commute but instead

$$\mathbf{AxB} = -(\mathbf{BxA})$$

Readers should prove this statement to themselves. In general, vector and matrix multiplication is not commutative. Scalar products are therefore a special case since these operations do commute.

Geometrically, the cross product can be understood as a vector that is perpendicular to both of the root vectors and whose magnitude is equal to the area of the parallelogram traced out by **A** and **B**. Writing the vector product in terms of components is best done by writing a determinant:

$$\begin{aligned} \mathrm{AxB} &= \begin{vmatrix} \mathbf{i} & \mathbf{j} & \mathbf{k} \\ A_x & A_y & A_z \\ B_x & B_y & B_z \end{vmatrix} \\ &= \mathbf{i}(\mathrm{A}_y\mathrm{B}_z - \mathrm{A}_z\mathrm{B}_y) - \mathbf{j}(\mathrm{A}_x\mathrm{B}_z - \mathrm{B}_x\mathrm{A}_z) + \mathbf{k}(\mathrm{A}_x\mathrm{B}_y - \mathrm{B}_x\mathrm{A}_y) \end{aligned}$$

Appendix B
Quantum Electrodynamics

The theory that describes the interaction of electromagnetic radiation with matter is called quantum electrodynamics or QED. It is a unification theory of photons and electronic distributions that incorporates electromagnetic theory (Maxwell's equations), relativity, and quantum mechanics. This will be a cartoon level presentation of the theory, but it should add some perspective to the overall picture we have drawn in this book. We know that light is made up of photons. Photons are massless and have a wave/particle duality. We are familiar with the uncertainty inherent in complementary observables, position and momentum, and energy and time. We have also talked extensively about the interaction of the electric force between particles and between the photonic electromagnetic wave. How does the electric force get across space to participate in these interactions?

Let us consider two particles with similar charge. We know that they repel one another because of the repulsive electric force. According to QED, the electric force is composed of photons that pass between the two particles. These are special photons called *virtual photons* and are a quantum mechanical phenomenon because they exist as a consequence of the uncertainty of time-energy. Virtual photons are created out of "nowhere" and, during their existence, the energy of the charged particle–virtual photon system is greater than before the photon existed. This post-production energy is equal to the energy of the photon and the particle. The energy necessary to make these virtual photons unbalances the conservation of energy ledger but only for a limited time. A large energy deficit can only exist for a very short time but a smaller energy deficit can be tolerated for a longer time. The limit on this behavior is $\Delta E\ \Delta t \geq h$. So a charged particle is a particle surrounded by a cloud of virtual photons constantly being created, ejected, and absorbed. Charge is therefore the ability to generate virtual photons and the electric field is the virtual photon cloud.

How does this description account for the known behaviors of charges and the electric force? Namely

(1) A repulsive force between two like charged particles,
(2) Increasing force with greater charge,
(3) A force that diminishes with distance,
(4) A force that decreases to zero only at infinity,
(5) The production of electromagnetic radiation (real photons).

Consider this picture: Our charged particle is constantly creating and spitting out virtual photons. If two particles are near to one another, they can exchange virtual photons. The repulsive force can be appreciated by an analogy from classical mechanics. These photons have momentum. When a particle ejects a photon, it recoils just as a gun does when fired. The particle catching the photon absorbs the momentum and it also is knocked back. Thus, the exchange of the virtual photons leads to the concerted motion (acceleration) of the two particles away from one another – a repulsive force.

These photons come in a variety of energy sizes. A large energy virtual photon may be ejected but has a very short lifetime because of indeterminacy. Even at the speed of light, it cannot travel very far away from the charged particle before it must return to pay back its "borrowed" energy. Alternatively, a very low-energy virtual photon has a longer lifetime and can travel much further away from the particle. Thus, the high-energy photons are only found near the charged particle and only low-energy photons are found at a distance away – the force is weaker with distance. A much more detailed mathematical analysis could show that the relationship to be the familiar inverse square law. A larger charge produces more virtual photons so the force will be proportionately larger.

Virtual photons can be made into real photons by a process such as a collision that knocks the charge away from its cloud of photons. The isolated virtual photons cannot be reabsorbed by the absent particle and the energy of the collision that removed the particle is available to the photons. They become real using the energy of the interaction and radiate. Thus, electromagnetic waves are generated from virtual photons when a charged particle is accelerated, shaking off a portion of the virtual photon cloud. The description of the quantum electrodynamic interaction accounts for the classical and quantum effects that we have been alluding to in the interactions between light and matter.

Appendix C
The Pre-Socratic Roots of Modern Science

While understanding of the nature of the world has probably been an essential quest of human beings at least since the Neanderthals 40 to 100,000 years ago practiced religious rituals in the burying of their dead, the general solution was a belief in either magical powers or gods and goddesses imbued with supernatural and often unpredictable powers over the natural world and its human inhabitants. The pre-Socratic Greek philosophers were the first thinkers to reject explanations of natural events as the result of actions by supernatural beings. The first true philosophers were the members of the Milesian school, Thales of Miletus, and his students, Anaximander and Anaximenes. These thinkers were the first to propose an understanding of the cosmos by reason or *logos* rather than by myth. The Milesian school explained natural phenomena by reference only to observable events. The primary goal of these investigations was to rationally comprehend the cosmos in terms of the "principle" or the fundamental substance from which the natural world was made. The Milesians were among the first to formulate a scientific philosophy and they concerned themselves with a philosophy of the *physis* and not with a philosophy that concerned the specific nature of man. Thales reasoned that this substance was water because of the observations that it was wholly from water that life came and into which life dissolves. Anaximander reasoned further that a more ephemeral "principle", the *apeiron*, meaning with unlimited internal and external dimension, was the fundamental substance of the cosmos. He argued based on his observations of nature that the Earth was suspended in an equilibrium of forces (rather than floating in a pool of water as Thales thought) and that life evolved from organisms originally found in the water and that grew more complex as they adapted to their environment on the land. In like fashion, Anaximenes argued for air as the "principle" because of his observations on the heating and cooling effect of air being forced from the lips at varying pressures. It is the combination of empirical observation as the basis for the application of logic in an effort to understand the cosmos that makes these early philosophers the founders of a modern-form philosophy of science.

Though in many ways the pre-Socratic philosophers seem familiar to us it is a mistake to think that their development of scientific observation was with fully modern intent. Perhaps this is best appreciated by considering the Pythagoreans who were akin to a religious cult whose principal aim was to achieve a particular lifestyle, and scientific inquiry was the *means* to this goal. The Pythagoreans altered the Milesian or Ionian idea of the “principle” and argued that the “principle” was *number* and the *constitutive elements of number*. The Pythagoreans believed that through systematic quantification of the natural world they would be able to extend the natural mathematical relationships that they were discovering in music, seasonal time, and life cycles to include the very workings of the unseen cosmos. Thus for the first time the Universe was seen as an ordered whole no longer dominated by mysterious or obscure forces. Through the rule of *number,* order is found, and the rule of rationality and truth becomes transparent to the mind. The Pythagorean concept of *number* is different from our use of numbers as a symbolic tool through which we formally map a representation of the natural world. To the Pythogoreans, *number* was a real thing and in fact it was the most real irreducible of things and as such could be used to constitute other things. Numbers thereby were tangible and not symbolic and occupied space thus having an arithro-geometry. The Pythagoreans, however, were themselves not pure cold rationalists since their ultimate goal was to replace the sacred religious mysteries with the mysteries of science and thus achieve an ecstasy of the mind. This state is to be found by the “contemplative life” in which purification is found in contemplation of the truth which is found via knowledge.

Thus the usefulness of true scientific and mathematical inquiry was established by the Ionians and Pythagoreans for the greater purpose of finding the order in the Cosmos beyond that given by religion and mysticism. But the use of the knowledge gained was to find truth and a way of living that would be both personally and socially successful. The trend in the history of the Greek philosophers thus was to move from a philosophy of the natural world to a moral philosophy of man. Placing of man at the center of the debate was necessary since the specific actions or events involving an individual could not be usefully related to a cosmic philosophy. Thus to determine the “principle” governing the actions of an individual in humankind became the goal of the moral philosophers who sought to understand the nature or essence of man. The philosophical description of the difference between a philosophy of the cosmos (*physis*) and a moral philosophy of man has a strong resonance with the modern philosophical debates triggered by quantum mechanics against a tradition of classical/deterministic mechanics and thermodynamics. In fact the progression from the Ionian scientists and Pythagorean mystical mathematicians to the humanistic Sophists may well presage the debates between Einstein and Bohr which philosophically can be reduced to the epistemological issue of whether the state of the system depends on the question asked by the observer (invariably directed by the knower) [Bohr] or is independent of the observer [Einstein].

Appendix D
The Poisson Function

The Poisson function:

$$P(\text{outcome is } k) = \frac{e^{-\lambda}\lambda^k}{k!} \tag{D.1}$$

can be shown to be a legitimate function because the sum of the probabilities given in D.1 is 1.

$$\sum_{n=0}^{\infty} \frac{e^{-\lambda}\lambda^k}{k!} = e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^k}{k!} = 1 \tag{D.2}$$

The term $\sum_{n=0}^{\infty} \frac{\lambda^k}{k!}$ can be called e^{λ} and is identical in form to the standard series:

$$e^x = \sum_{n=0}^{\infty} \frac{\lambda^k}{k!} = 1 + x + \frac{x^2}{2!} + \frac{x^3}{3!} + \ + \cdots \tag{D.3}$$

in similar fashion e^{-x} is in a form given by a standard power series:

$$e^{-x} = \sum_{n=0}^{\infty} \frac{\lambda^k}{k!} = 1 - x + \frac{x^2}{2!} - \frac{x^3}{3!} + \frac{x^4}{4!} - \cdots \tag{D.4}$$

and the product:

$$e^x e^{-x} = 1 \tag{D.5}$$

This is the result that we are seeking.

Appendix E
Assumptions of a Theoretical Treatment of the Ideal Gas Law

The theory for ideal gases makes the following assumptions:

The gas particles

- Are very small and spherical.
- All have non-zero mass.
- Are numerous so that statistical treatments apply.
- Are in constant random motion.
- Have negligible interactions except on collision

Collisions:

- With the wall are elastic.
- With the wall are constantly occurring.
- With other gas particles are elastic.

The volume of the gas compared to the volume of the container is negligible.
The average kinetic energy depends only on the temperature of the system.
Relativistic and quantum mechanical effects are negligible.
The time during collision of molecule with the container's wall is negligible compared to the time between successive collisions.
The equations of motion of the molecules are time reversible.

Appendix F
The Determination of the Field from the Potential in Cartesian Coordinates

The displacement vector $d\boldsymbol{s}$ can be decomposed into Cartesian coordinates,

$$d\boldsymbol{s} = dx\,\mathbf{i} + dy\,\mathbf{j} + dz\,\mathbf{k} \tag{F.1}$$

recalling that **i**, **j**, and **k** are unit vectors.

The change in the potential can be written as

$$dV = -\mathbf{E}\,d\boldsymbol{s} = (-\mathbf{E}_x\,dx) - (-\mathbf{E}_y\,dy) - (-\mathbf{E}_z\,dz) \tag{F.2}$$

The potential must generally be considered to depend on all three space coordinates and therefore moving from r to $r + d\boldsymbol{s}$:

$$x\mathbf{i} + y\mathbf{j} + z\mathbf{k}\ \mathit{to}\ (x + dx)\,\mathbf{i} + (y + dy)\,\mathbf{j} + (z + dz)\,\mathbf{k} \tag{F.3}$$

We must take this spatial dependency into account. This is accomplished by using partial derivatives, which gives the following result:

$$dV = \frac{\partial V}{\partial x}dx + \frac{\partial V}{\partial y}dy + \frac{\partial V}{\partial z}dz \tag{F.4}$$

Equation (F.2) can now be written in terms of the coefficients of dx, dy, and dz:

$$E_x = -\frac{\partial V}{\partial x}, E_y = -\frac{\partial V}{\partial y},\ \text{and } E_z = -\frac{\partial V}{\partial z} \tag{F.5}$$

which gives the electric field vector in terms of the derivatives of the electric potential:

$$\mathbf{E} = -\frac{\partial V}{\partial x}\,\mathbf{i} - \frac{\partial V}{\partial y}\,\mathbf{j} - \frac{\partial V}{\partial z}\,\mathbf{k} \tag{F.6}$$

Appendix G
Geometrical Optics

G.1 Reflection and Refraction of Light

In a practical sense, geometric optics is the language of optical instrument design. It has important application in describing the behavior of mirrors, prisms, and lenses. It is useful to understand these aspects of microscopes, spectrophotometers, and similar optical devices used in the laboratory. The entire field of geometrical optics can be derived from two laws, *the law of reflection* and *the law of refraction*. From an experimental standpoint, we know that when a light ray strikes a boundary between two transparent substances in which the velocity of light is significantly different, the ray of light will divide into a reflected and refracted ray. (The relationship between wave impulses be noted.) We imagine (Fig. G.1) a ray (the incident ray), falling on a boundary so that it makes an angle, θ, with a line drawn normal or perpendicular to the surface of the boundary. θ is called the *angle of incidence* and the plane described by the incident ray and the normal is the plane of incidence. The law of reflection is

> The reflected ray lies in the plane of incidence and the angle of reflection is equal to the angle of incidence, $\theta = \theta^{'}$

> The law of refraction, often known as Snell's law, is

> The refracted ray lies in the plane of incidence. The sine of the angle of refraction bears a constant ratio to the angle of refraction.

The ratio in this law, $\frac{\sin\theta^1}{\sin\theta^2}$ equals a constant which relates the refractive indices n of the two transparent media. If the left-handed media in Fig. G.1 is a vacuum (or for all practical purposes air), which is defined to have a refractive index of 1, then the constant is the refractive index of the substance on the right side. Snell's law is commonly written as

$$\frac{\sin\theta^1}{\sin\theta^2} = \frac{n^1}{n^2} \text{ or}$$

$$n_1 \sin\theta_1 = n_2 \sin\theta_2 \tag{G.1}$$

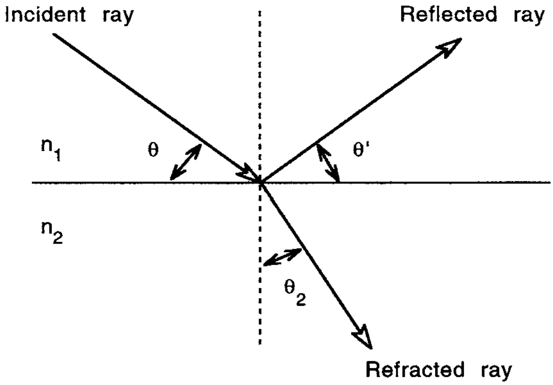


Fig. G.1 The laws of reflection and refraction are central to the application of geometrical optics

A very convenient approximation for working with small angles is to let the sines of the angles equal the angles:

$$\frac{\theta^1}{\theta^2} = \frac{n^1}{n^2} \tag{G.2}$$

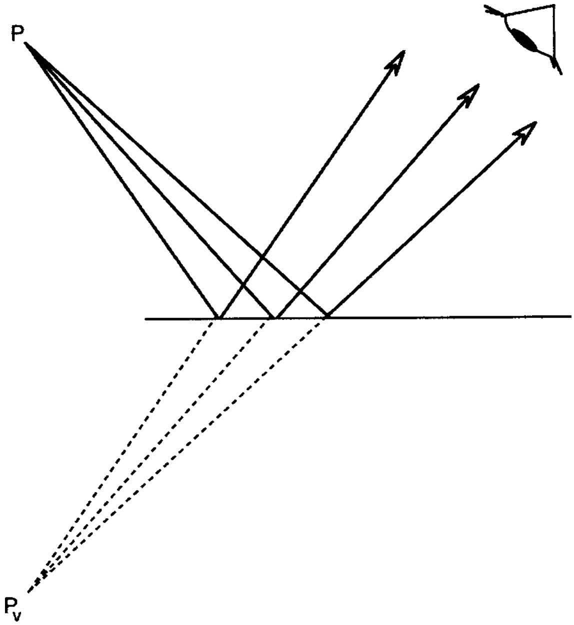
A link between the law of reflection and refraction can be made if a reflected ray is considered to be transmitted from the first medium with a refractive index of n_1 back into itself ($n_1 = n_2$), then it is easy to see that under those conditions Eq. (B.2) is equivalent to the law of reflection.

G.2 Mirrors

G.2.1 The Plane Mirror

A mirror can be considered as a medium of infinite weight. When struck by the impulse the wave is completely reflected. No energy is transmitted through the mirror (i.e., no refraction takes place). Thus we are only concerned with the law of reflection. A point source (P) emits a series of light rays that strike a plane mirror and are reflected from the mirror surface (Fig. G.2). The pattern of the reflected rays diverging from the mirror surface is exactly the same as if the rays had originated from a point source (P_v) behind the mirror. To an observer above the plane of the mirror, the image of P appears to be originating from the point source P_v . The reflected image of P cannot be distinguished from a theoretical or virtual image originating at P_v. This image is called a *virtual image*. A geometrical construction using the law of reflection will demonstrate that the image point is symmetrical to the object point with the plane of the mirror equidistant and perpendicular to a line drawn between the two points. The image can be seen by an observing eye at a

Fig. G.2 The plane mirror



variety of angles as shown and the object does not need to be in front of the mirror. Plane mirrors do not magnify. The image will be the same size as the object.

G.2.2 The Concave Mirror

The case is different with a concave spherical mirror surface. A group of rays from a source (P) lying on the axis are reflected from the mirror. In this situation the reflected rays are found to converge at a point P_C and then to diverge from P_C as if the object were actually there. An image perceived by observing the diverging rays from P_C is called a *real image* because the light actually does emanate from the image point. If a sensing device (an eye, electronic or photographic plate or ground glass screen, etc.) were placed at P_C, the image would be demonstrated. This is in distinction to placing a similar device at "P_V" where no image would be formed because there is no light to be found there. The sensing device that is placed in the path of the diverging rays, however, will not be able to differentiate between the virtual or real images because at that point they are indistinguishable as to source.

An important observation about spherical mirrors is that only the *paraxial rays* are reflected through P_C. Paraxial rays are those rays whose angles of reflection (or refraction) are small enough that the cosines are equal to unity and the sines are equal to the angles. Other rays converge to points near P_C but the lack of coregistration at P_C causes a blurred image and this effect is named *spherical aberration*. If we consider only paraxial rays, the relationship between the object distance, s, the

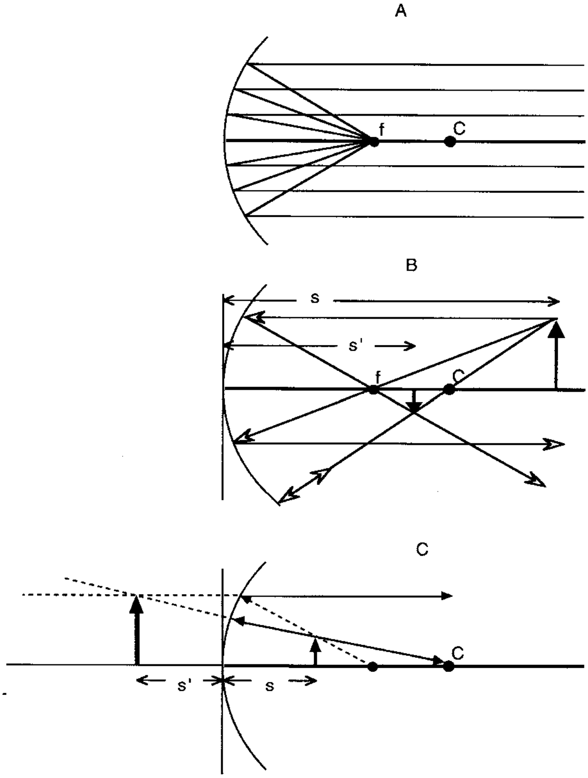


Fig. G.3 The concave mirror. In this figure, f is the focal point, C is the radius of curvature, s and s′ are the object and image distances, respectively

image distance, s', and the radius of curvature of the mirror, r, can be related by

$$\frac{1}{s} + \frac{1}{s'} = \frac{2}{r} \tag{G.3}$$

If the object distance is significantly greater than the curvature of radius, then the $\frac{1}{s}$ term can be ignored and the image distance or the focal length of the mirror, f, can be found as

$$f = \frac{1}{2}r \tag{G.4}$$

Equation (G.3) can be written in terms of the focal length:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{G.5}$$

The focal point, which is located at f, is the point at which parallel rays corresponding to plane waves coming from an infinite distance are focused. Thus a spherical mirror can be used to convert parallel plane waves into spherical waves that converge at the focal point. Because of the wave property of reversibility, if a source of spherical waves is placed at the focal point and aimed at the axial line, it will generate parallel plane waves. Thus using a point source of light, a spherical mirror and some method of preventing the light from falling on the non-paraxial portions of the mirror, a collimated beam of light can be generated.

The image formed by a concave mirror can be located by geometrical means as shown in Fig. G.3. Note that because we are limiting our analysis to paraxial rays, the plane of reflection is the *tangent plane* to the mirror. Real images will be inverted and of different size than the object except at the point of unit magnification where $s = C$, the center of curvature. The magnification of the mirror (and in fact of any optical system) is the ratio of s' to s.

$$m = \frac{-s'}{s} \tag{G.6}$$

The conventions adopted for this equation are related to the condition that real images are formed on the mirror side and that virtual images are formed behind the mirror. Thus all distances on the real side are positive and all distances on the virtual side are negative. For concave mirrors s and r are positive and the sign of s' depends on whether the image is real or virtual. The convention of the negative magnification tells us that the image is inverted.

If the object is beyond the center of curvature, the image formed will be real, inverted, and smaller in size than the object. When the object is inside the focal point, the image is virtual, upright, and larger than the object.

G.3 Image Formation by Refraction

The rays of light that are transmitted through the transparent interface are refracted consistent with Snell's law.

The scheme for describing image formation by refraction is shown in Fig. G.4. In general, the conventions used in discussing refracting surfaces are similar to those involving mirrors. Thus we can describe the following:

(1) Only paraxial rays are considered;
(2) The real image is formed to the right of the surface when the object is to the left;
(3) Virtual images will be to the left (again if the object is left);
(4) s' and r will be positive if the image and center of curvature lie to the right of the surface.

A geometrical treatment similar to that above for the mirror but using Snell's law shows the relationship between image and object distances, their refractive indices

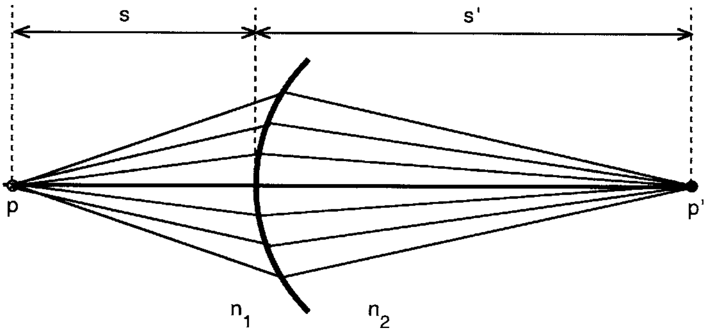


Fig. G.4 General diagram showing the conventions used in describing refracting systems

and the radius of curvature (r):

$$\frac{n^1}{s} + \frac{n^2}{s'} = \frac{n^1 - n^2}{r} \tag{G.7}$$

We can use this formula to determine image position for lenses and for refracting media with a flat surface, such as an fish tank. In the case of a lens, two surfaces are generally closely approximated and the position of the image formed by a lens is found by considering the refraction of each surface separately. In general, if presented with this problem, the best plan of attack is to simply determine the new focus by the first surface and then use the new surface's radius of curvature and refractive index and recalculate. For most biophysical scientists, the few times in a lifetime that several surfaces will have to be chased through will be so rare that approaching the problem systematically is the simplest solution. The one special case, which is also probably the most common is the case of the thin lens. Most lenses are not thick enough to actually form the first image before light is refracted at the second surface and in fact the thickness of the lens can effectively be ignored. The focal length of a thin lens is defined to be the image distance when the object distance is very large, essentially at infinity. In the formula for the focal length of a thin lens, $s = \infty$ and the focal point f is the image distance s':

$$\frac{1}{f} = (n - 1)\left(\frac{1}{r^1} - \frac{1}{r^2}\right) \tag{G.8}$$

or in terms of a formula called the *thin lens equation* which is the same as the mirror equation:

$$\frac{1}{s} + \frac{1}{s'} = \frac{1}{f} \tag{G.9}$$

A light ray passing through a double convex lens will be bent toward the axis of the lens by both surfaces. Such lenses are called positive or converging lenses. If parallel rays are made to shine through this lens from the left (as if the light source was placed at ∞) the rays will be focused at a distance f to the right of the lens at a focal point generally designated F' (called the second focal point). A similar procedure with light shining right to left will locate F (the first focal point). A light source diverging from either focal point will leave the lens as parallel rays. The size of the image will depend on the placement of the object relative to the focal. The magnification is given by the same formula as used for mirrors. In similar fashion as we did for mirror images, the images formed by a lens can be found by using a ray diagram.

In the thin lens approximation, the rays are treated as if they all bend at one plane passing through the center of the lens. In most optical instruments more than one thin lens is used and the overall behavior of the instrument can be described by two principal planes that often coincide fairly closely with the surfaces of the first and last lenses. Principal plane analysis is appropriate if the light starts and ends in the same medium (such as air) and can be summarized,

(1) Parallel light coming from the left will be focused on the right at a distance form the second principal plane equal to the focal length.
(2) Parallel light from the right will be focused at a distance equal to the focal length from the first principal plane.

A thin lens then is just a special case of this general formulation in which the two principal planes coincide. The geometry of the light rays obey the relation $aa' = ff$' with the magnification given by $-\frac{f}{a}$. The meaning of the negative sign is that the image is inverted in the image plane with respect to the object plane. These ideas are used in Chapter 29 where the formation of images in a compound microscope is discussed.

G.4 Prisms and Total Internal Reflection

Our discussion of geometrical optics will end with the results of two important effects of refraction which will also lead us into physical optics. The refractive index is an empirically measured quantity that actually carries a great deal of information about molecular structure of the refracting medium. To the chemist, this makes the understanding of the physics of light interaction a window into the chemical nature of the material. We will use such an analysis to help us gain a perspective on molecular properties.

If a light ray passes from a medium of higher refractive index to a medium of lesser n, depending on the angle of incidence, light will be refracted or reflected. If the angle of incidence is severe enough, all of the light will be reflected and the light will stay inside the denser material. This is called *total internal reflection*. Total internal reflection occurs when the incident light is above a certain critical angle, θ_{c}. Refraction occurs below θ_{c}. The critical angle depends on the ratio of the refractive indices:

$$\sin\theta_c = \frac{n_2}{n_1} \tag{G.10}$$

For glass and air it is approximately 41° and so an angle of incidence greater than 42° will lead to reflection. This is why a 45-45-90 prism is an effective mirror and is used in many optical instruments such as microscopes, binoculars, beam splitters, and especially in applications where corrosion of a silvered mirror surface might be problematic. Total internal reflection is also the principle by which fiber optic devices can carry light or, if the fibers are topographically oriented, an image.

What happens to the light that escapes from a prism? As previously mentioned, the refractive index is slightly dependent on wavelength:

$$n \propto \frac{1}{\lambda} \tag{G.11}$$

Therefore when light is refracted, the blue components are bent more than the red components. In prism the light is refracted twice with each surface adding to the separation or dispersion of the beam of light. Rainbows are formed because of dispersion and total internal reflection. When sunlight strikes the water droplets, it is reflected from the side of the droplet opposite the viewer (the critical angle of water/air is approximately 40°). This reflected light then is refracted and dispersed as it exits the drop. The rainbow appears as an arc when viewed from the ground because the drops at an angle of 40° can only exist above the ground. This is why circular rainbows can be seen from airplanes.

Appendix H
The Compton Effect

Particles must possess linear momentum. If they collide with other particles, momentum will be transferred and scattering must occur. If light is composed of photons and photons are particles, then if a beam of light falls on a stream of electrons, the light should be scattered with a change in momentum that is dependent on the angle of interaction. A change in the momentum or kinetic energy of light will appear as a change in frequency of the scattered beam. [$E_{\text{photon}} = h\nu$ and $m = 0$; $p = \frac{h\nu}{c}$ because $E_{\text{relativistic}} = (m^2c^4 + p^2c^2)^{\frac{1}{2}}$.] A. H. Compton performed this experiment in 1922 using a beam of gamma ray light that interacted with a stream of free electrons. He demonstrated that for each angle of incidence, a single wavelength shift of the gamma light beam was found as given by

$$\lambda_{\text{final}} - \lambda_{\text{initial}} = \left(\frac{2\,h}{m_e c}\right) \sin^2 \frac{1}{2}\theta \tag{H.1}$$

where m_{e} = electron mass. These results support the model of particles interacting with “transfers of energy occurring in single encounters” and do not support the interaction of a gradual transfer of energy as required by the classical behavior of waves.

Appendix I
Hamilton's Principle of Least Action/Fermat's Principle of Least Time

We talk of the shape of a trajectory of a particle moving in space. Whether a cannon ball or a water molecule, the path the particle takes is a unique one. There are in fact very many available paths but the path taken is unique. If at every point along the observed path the kinetic energy minus the potential energy of the particle is integrated over the time of the entire path, the observed path will have the smallest possible measure of this quantity called the *action*. Any other path will have a larger action. The action, *S*, for those who wish a more mathematical expression is an integral over the time of the path of another function, the Lagrangian, *L*. The Lagrangian is a function of the position and velocity of a particle, i.e., its potential and kinetic energy:

$$L(x,x') = T - U = \frac{1}{2}mv^2 - U(x) \tag{I.1}$$

$$S = \int_{t_1}^{t_2} L\left(x,x'\right)dt \tag{I.2}$$

The fact that particles are observed to move in these least action paths was proposed by W.R. Hamilton in the 19th century and was shown to be identical to Newton's laws of motion. In a medium of constant potential, the *principle of least action* dictates that particles travel in straight lines. The other energy-conveying object is the wave, and waves also travel in straight lines in a uniform medium. In this way the principle of least action suggests that it is similar to Fermat's principle of least time and furthermore that particulate matter has a wavelike character! Hamilton did not discover the quantum theory, but de Broglie saw the relationship between Hamilton's and Fermat's principles and did substantially advance the theory.

Fermat proposed his *principle of least time* in 1650. Fermat stated that of all the possible paths a ray of light could take going from point to point the preferred path would be the one that took the least time. This principle explains the straight-line behavior of light in a single medium and the observed reflection and refraction seen with mirrors and lenses. The rules of refraction are derived from the different speed of propagation of light in different media. If the speed varies as light travels on a path from one medium to another, the light will bend to the path which leads to the shortest total time. The reason that the path of least time is the one that is observed derives from the fact that when a wave travels a longer distance, the phase of the arriving wave is different depending how long it took that particular wave to get there. All of the neighboring waves which travel a long distance are out of phase with one another and destructively interfere when they arrive at the final point. The overall superposition of all of these waves leads to no net wave arriving except from the shortest distance because all of the other neighbors annihilate each other.

Appendix J
Derivation of the Energy of Interaction Between Two Ions

The interactional energy between two ions can be found simply. The starting point is that there is an electrostatic interaction between the two ions given by Coulomb's law:

$$F = \frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon r^2} \tag{J.1}$$

The energy of the interaction is found by calculating the electrical work necessary to bring a total charge q_1 in infinitesimal increments dr, until it rests a distance r from charge q_2. This work is calculated from the equation that defines work:

$$w = U_{i-i} = \text{Force} \times \text{Distance} \tag{J.2}$$

$$\int dw = -\int\limits_{\infty}^{r} \frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon r^2} dr \tag{J.3}$$

The constants can be removed from the integral:

$$\int dw = -\frac{1}{4\pi \varepsilon_0 \varepsilon} \int\limits_{\infty}^{r} \frac{q_1 q_2}{r^2} dr \tag{J.4}$$

which gives the result:

$$U_{\text{i - i}} = \frac{q_1 q_2}{4\pi \varepsilon_0 \varepsilon_{\text{r}}} \tag{J.5}$$

which is the result of interest. When two oppositely charged particles are brought together, the sign of the interactional energy will be negative, indicating that there is an attractive force between them.

Appendix K
Derivation of the Statement, $q_{\mathrm{rev}} > q_{\mathrm{irrev}}$

In Chapter 11, it was demonstrated that for a process producing work, in the surroundings the maximal work can be obtained from a reversible process only, allowing the following to be written:

$$w_{\mathrm{rev}} > w_{\mathrm{irrev}} \tag{K.1}$$

Consider a change in state carried out first reversibly, and then subsequently irreversibly. By the first law:

$$\Delta U = q_{\mathrm{rev}} + w_{\mathrm{rev}} \text{ and } \Delta U = q_{\mathrm{irrev}} + w_{\mathrm{irrev}} \tag{K.2}$$

However, since work is done on the surroundings, w_{rev} and w_{irrev} are given a negative sign, hence

$$\Delta U = q_{\mathrm{rev}} + (-w_{\mathrm{rev}}) \text{ and } \Delta U = q_{\mathrm{irrev}} + (-w_{\mathrm{irrev}}) \tag{K.3}$$

In both cases, the change in state is identical, so these equations can be equated:

$$q_{\mathrm{rev}} - w_{\mathrm{rev}} = q_{\mathrm{irrev}} - w_{\mathrm{irrev}} \tag{K.4}$$

which is the same as writing:

$$q_{\mathrm{rev}} - q_{\mathrm{irrev}} = w_{\mathrm{rev}} - w_{\mathrm{irrev}} \tag{K.5}$$

But

$$w_{\mathrm{rev}} > w_{\mathrm{irrev}} \tag{K.1}$$

so

$$q_{\mathrm{rev}} - q_{\mathrm{irrev}} > 0 \tag{K.6}$$

This result leads to

$$q_{\text{rev}} > q_{\text{irrev}} \tag{K.7}$$

which is the solution used in Eq. (12.6).

Appendix L
Derivation of the Clausius–Clapeyron Equation

The relationship between pressure and the boiling point of a pure solvent can be simply derived by considering the equilibrium constant for the process:

$$X_{\text{(liquid)}} \rightarrow X_{\text{(gas)}} \tag{L.1}$$

The activity of the pure liquid is unity, and if the gas is dilute enough to be considered an ideal gas, its activity is simply its pressure. *K* may be written in this case as

$$K = \frac{a_{\text{(gas)}}}{a_{\text{(liquid)}}} = \frac{P}{1} \tag{L.2}$$

The temperature dependence of the equilibrium constant can be given by

$$\ln \frac{K_2}{K_1} = \frac{-\Delta H}{R} \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \tag{L.3}$$

Combining these last two equations gives the Clausius–Clapeyron equation:

$$\ln \frac{P_2}{P_1} = \frac{-\Delta H_{\text{vaporization}}}{R} \left(\frac{1}{T_2} - \frac{1}{T_1} \right) \tag{L.4}$$

Appendix M
Derivation of the van't Hoff Equation for Osmotic Pressure

At equilibrium, there will be no net flow of solvent through the membrane. If one phase is the solvent in standard state, μ^{o}, and the other phase, μ, is the solvent, A, with the added solute, b, and the added external pressure is $+\pi$, the following can be written as

$$\mu^{0}\left(T,P\right)=\mu\left(T,P,+\pi\right) \tag{M.1}$$

This can be rewritten as

$$\mu^{\mathrm{o}}\left(T,P\right)=\mu^{\mathrm{o}}\left(T,P,+\pi\right)+RT\ln a_{A} \tag{M.2}$$

The solute that is added to make the solution has decreased the activity of the solvent by the amount $RT\ln a_A$:

$$\mu^{\mathrm{o}}\left(T,P\right)-\mu^{\mathrm{o}}(T,P,+\pi)=RT\ln a_{A}=\Delta\mu \tag{M.3}$$

Therefore, the problem now is to determine how much external pressure must be added to the solution side to raise the activity of the solvent by the amount $RT\ln a_A$. At constant temperature, the fundamental equation reduces to

$$\Delta G=\Delta\mu=\int_{P_1}^{P_2}\overline{V_A}dP=\overline{V_A}\left(P_2-P_1\right) \tag{M.4}$$

$\overline{V}$ is the molar volume of the solvent. Since P_2-P_1 is equal to $-\pi$, Eqs. (M.3) and (M.4) can be combined to give

$$\Delta\mu=RT\ln a_{A}=\overline{V_A}\pi \tag{M.5}$$

This allows direct evaluation of the activity of the mole fraction of the solvent from the osmotic pressure:

$$\ln a_{A}=-\frac{\overline{V_A}\pi}{RT} \tag{M.6}$$

For a dilute solution, $\ln a_A$ can be approximated by writing first the activity in terms of the mole fraction of A:

$$\ln a_A = \ln X_A \tag{M.7}$$

but $\ln X_A$ can be written in terms of the mole fraction of the solute, X_b:

$$\ln X_A = \ln (1 - X_b) \tag{M.8}$$

For a dilute solution, $X_b = 1$, and therefore the logarithm can be expanded and all but the first term ignored:

$$\ln (1 - X_b) \approx -X_b \approx \frac{-n_b}{n_A + n_b} \approx \frac{-n_b}{n_A} \tag{M.9}$$

since $n_b = n_a$. Substituting this result into Eq. (M.6) gives the following result:

$$\pi = \frac{n_b RT}{n_a \overline{V}_A} \tag{M.10}$$

The volume of the solute is very small and can be ignored, so the total volume, V, can reasonably be considered equal to the solvent molar volume, $n_a \overline{V}_A$. This gives the result sought:

$$\pi = \frac{n_b RT}{V} \text{ or } \pi = cRT \tag{M.11}$$

the van’t Hoff equation for osmotic pressure.

Appendix N
Fictitious and Pseudoforces – The Centrifugal Force

We will be concerned with several kinds of pseudoforces in this book. Pseudoforces appear as forces in a system and are best regarded as properties of the system rather than the entities affected. Diffusion, entropy, and centrifugal force are all examples of forces derived from system interactions. Here we consider some of the more interesting aspects of centrifugal force.

Centripetal forces are real forces in contrast to the centrifugal force which is actually a *fictitious force*. Yet we certainly take advantage of the centrifugal force. After all we use centrifuges not “centripetal-fuges”. The reason that the centrifugal force is fictitious is that it involves a non-inertial frame of reference. A system moving in a circle even at constant angular velocity is constantly accelerating thus is not in an inertial frame of reference. An observer or a reporter system, A, within a system that is accelerating (we make it constant acceleration for convenience) will believe themselves to be at rest, but to an observer outside the accelerating system, the observer A will be seen as accelerating. An example will help to understand these fictitious forces: A box with a cable attached is placed in space so that no force is present at rest. You are standing inside the box on the surface opposite the wall to which the cable is attached. A weighing scale is next to you. Now the elevator cage is moved through space, pulled by the cable at a constant acceleration of 9.8 m/s^2. The force moving the cage is the tension on the cable and the acceleration vector is pointed up from your feet toward your head. But what do you perceive? Before the box started to move you were a mass at rest. Once the box starts to move you still try to remain a mass at rest but now the floor is accelerating toward your resting body at 9.8 m/s^2. If you try to leap away from the floor, the floor will keep coming toward you and your *perception* will be that you are moving toward the floor! If you stop trying to jump out of the way of the floor and instead step onto the scale you will find that the scale indicates that you weigh exactly your weight on Earth. Of course when you jumped into space, you perceived that you were falling to the floor as well, just like on Earth. Because you perceive your system to be at rest and make interpretations as if the system were an inertial frame of reference you will interpret that a force just like gravity exists and is pulling you toward the floor. Of course for those of us outside your frame of reference, watching you leaping up so that the accelerating floor can hit you in the face, you look pretty silly. We know that the only force in the system is the tension on the cable and that you are constantly accelerating.

Your explanation for moving toward the “floor” as falling is wrong and is attributed incorrectly by you to a fictitious force. An example like this was used by Einstein to show that gravity and an accelerating frame of reference cannot be distinguished by an observer in the non-inertial system.

With this example as background, centrifugal force can be a little more easily understood. Consider a rider on a carousel. To the observer standing away from the carousel, the rider sitting on a painted horse is accelerating centripetally. A centripetally directed normal force from the horse pushing on the rider is responsible for keeping the rider moving in a curved path. The rider, however, has the view that she is at rest and feels the normal force from the horse on her legs and arms. The conclusion drawn is that there is a centrifugal force that forces her against the horse balancing the inwardly directed normal force. As the rider moves from horse to horse, the force can be appreciated to vary with the speed and radius of the motion of the carousel.

Appendix O
Derivation of the Work to Charge and Discharge a Rigid Sphere

The work necessary to charge a rigid sphere is found simply by calculating the work required to bring an amount of charge *q* from an infinite distance to reside on the sphere. This is done by bringing infinitesimally small quantities of charge *dq* to the sphere until the total charge is reached. The work is found by the product of the charge and the electrostatic potential:

$$dw = \psi_r \, dq \tag{O.1}$$

Starting with a sphere of zero charge and adding infinitesimal amounts of charge until the charge $z_i e_\mathrm{o}$ is reached is accomplished by integrating:

$$w = \int dw = \int_0^{z_i e_\mathrm{o}} \psi_r \, dq \tag{O.2}$$

The electrostatic potential is given by

$$\psi_r = \frac{q}{r} \tag{O.3}$$

Substituting this into Eq. (O.2) gives:

$$\int dw = \int_\mathrm{o}^{z_i e_\mathrm{o}} \frac{q}{r} dq = \left[\frac{q^2}{2r} \right]_\mathrm{o}^{z_i e_\mathrm{o}} \tag{O.4}$$

$$w = \frac{(z_i e_\mathrm{o})^2}{2r} \tag{O.5}$$

This is the work for charging a sphere in a vacuum and in the CGS system. In SI units and in any dielectric medium, Eq. (O.5) can be written as

$$w_c = \frac{(z_i e_\mathrm{o})^2}{8\pi \varepsilon \varepsilon_\mathrm{o} r} \tag{O.6}$$

The work of discharging is simply the opposite of the work of charging and is written as

$$w_{\mathrm{dis}} = -\frac{(z_i e_{\mathrm{o}})^2}{8\pi \varepsilon\varepsilon_{\mathrm{o}} r} \tag{O.7}$$

Appendix P
Review of Circuits and Electric Current

P.1 Current Density and Flux

Electric current is defined as the total charge that passes through a given cross section of area A, per unit time. Because charge is conserved, the shape or orientation is not important, only the area of the cross section. Current, like charge, is conserved, and this property allows us to systematically trace current flow in a path or circuit from higher to lower potential. The average current is given by $\Delta I = \Delta Q / \Delta t$ and an instantaneous form can be written in differential notation. The SI unit for current is the *ampere* and is equal to 1 coulomb per second. The ampere represents the total or net movement of charge, and there are many cases where the details of charge movement are important. When considering the details of charge movement we will use the *current density* **J**, which represents the rate of charge per unit area through an infinitesimal area. Current density thus depends on the local direction of flow and is a vector (amperes per square meter) in contrast to the scalar quantity of current. Current density is treated in a fashion similar to electric flux and depends on the orientation of the flux through the cross-sectional area, $d\mathbf{A}$:

$$dI = \mathbf{J} \bullet d\mathbf{A} = JdA\cos\theta \tag{P.1}$$

$$I = \int \mathbf{J}\, dA \tag{P.2}$$

As a fact of history, Benjamin Franklin thought that it was positive charge that flowed and gave rise to current, and so by convention, the arrow of current direction represents the flow of positive charge. This is not a vector but rather an indication of the potential difference between two relative points. This convention is still used today although it is the negatively charged electrons that are generally mobile in the solid state.

P.1.1 Ohm's Law

When an electrical potential field is imposed across a conducting material, the amount of current flow depends on both the potential and the *resistance* to flow of current. This is the relationship quantified in *Ohm's law*:

$$I = \frac{E}{R} \tag{P.3}$$

where I is the current expressed in amperes, E is the electrical potential field expressed in volts, and R is the resistance expressed in ohms (Ω). The resistance of a sample of conducting material is dependent on both the geometry and the intrinsic *resistivity* of the material to conduction of current. For a conductor of resistivity ρ, the resistance will increase as the length, l, of the current path increases and the resistance will fall as the cross-sectional area, A, of the conductor increases:

$$R = \rho \frac{l}{A} \tag{P.4}$$

The units of ρ are Ω m^{-1} or Ω cm^{-1}. Alternatively, the reciprocal of resistivity is a measure of the ease with which current flows in a conducting medium; it is quantitated as *conductivity* and is given the symbol κ. Conductivity or specific conductance can be expressed therefore as follows:

$$\kappa = \frac{1}{RA} \tag{P.5}$$

The units of κ are Ω^{-1} m^{-1} or Ω^{-1} cm^{-1}. Resistivity or conductivity is determined in a cubic volume element of 1 m^3 or 1 cm^3. The conductivity of a variety of materials is listed in Table P.1.

Table P.1 Conductivity values (κ) for selected materials at 298 K

Material	Conductivity (Ω^{-1} m^{-1})
Silver	6.33×10^7
Copper	5.80×10^7
KCl (0.1 M)	1.33×10^0
Acetic acid (0.1 M)	5.20×10^{-2}
Water	4.00×10^{-6}
Xylene	1.00×10^{-17}

It is convenient to regard the conductivity as the reciprocal of the resistivity and use the *conductance*, ***G***, to express the relationship between the current and voltage of a sample. Ohm's law may be rewritten as

$$I = GE \tag{P.6}$$

The unit of G is the mho, the reciprocal ohm (Ω^{-1}).

P.2 Circuits

An electrical circuit is constructed by connecting a source of electric energy such as a battery, capacitor, or generator (these are conventionally called sources of *electromotive force*, *emf*) through a series of elements such as resistors, capacitors, and wires. In a complete circuit the current flows from the emf source through the elements and back to the emf source. One of the simplest circuits is a battery and bulb in which current from the battery flows down the potential gradient of the battery through the resistance element in the bulb generating heat and light. The same principles that govern this simple circuit also apply to much more complicated circuits such as is shown in Fig. P.4.

P.2.1 Useful Legal Relations

We will now set down, without great explanation, some of the elements of electrical circuit making that are useful in the study of biophysical processes. Many of the biological electrophysiological and electrochemical models of interest have been constructed, at least on an elementary level, as electrical circuits. A fundamental familiarity with the rules of the game in terms of electrical circuitry is necessary for model making and is an essential skill for comprehension and appreciation of many modern biophysical ideas.

P.2.2 Kirchoff's Rules

Consider a circuit such as the battery–bulb circuit depicted in Fig. P.1. Each bulb is electrically equivalent to a resistor through which a voltage drop occurs. Because current and charge are conserved, the potential change in traversing the complete circuit must be zero. All of the current leaving the emf source, in this case the battery, must return to the battery, and the full potential difference measured at the terminals must be dissipated throughout the circuit. This is the central accounting rule for analyzing a circuit and is called *Kirchoff's second law* or *loop rule*: *the sum*

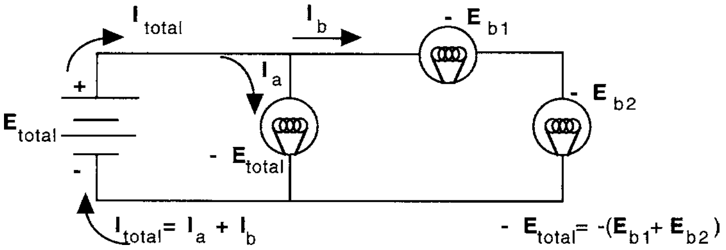


Fig. P.1 Simple battery and bulb circuit showing Kirchoff's relations

of potential changes around a closed loop is zero:

$$\sum_{\text{closed path}} \Delta V = 0 \tag{P.7}$$

Kirchoff's first law, also known as the *junction rule,* pertains to current flow and is applicable to single- and multiple-loop circuits: *the sum of the currents that enter a junction equals the sum of the currents that leave the junction*:

$$\sum I_{\text{in}} = \sum I_{\text{out}} \tag{P.8}$$

P.2.3 Capacitors in Series and Parallel

When capacitors are placed in a circuit in a parallel configuration, each plate takes the same potential as the conductor to which it is connected. Inspection of Fig. P.2 suggests that the effective area of the plates containing the charge increases proportionally when the capacitors are connected in parallel. Since capacitance is proportional to the area of the plates (Eq. P.9), the equivalent capacitance of a parallel circuit of capacitors will be

$$C_{\text{equiv}} = C_1 + C_2 + C_3 + \cdots + C_n \tag{P.9}$$

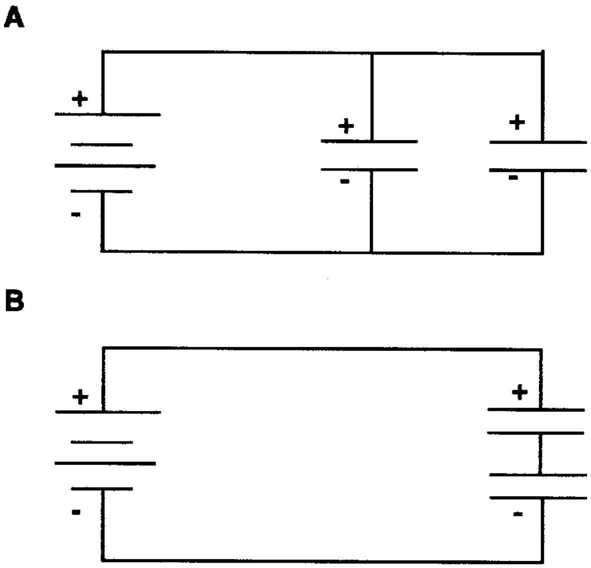


Fig. P.2 The equivalent circuit capacitance is dependent on the arrangement of the capacitor elements in the circuit branch

Alternatively when the capacitors are placed in a series arrangement, each capacitor will be found to have identical charges with each capacitor having a specific voltage drop associated with it. As we know from our accounting procedure given by Kirchoff's first law, $V = V_1 + V_2$ so we can write

$$V = V_1 + V_2 = \frac{Q}{C_1} + \frac{Q}{C_2} = Q\left(\frac{1}{C_1} + \frac{1}{C_2}\right) = \frac{Q}{C_{\text{equiv}}} \tag{P.10}$$

The relationship for the equivalent capacitance in a series circuit containing two capacitors is

$$C_{\text{equiv}} = \frac{1}{C_1} + \frac{1}{C_2} = \frac{C_1 C_2}{C_1 + C_2} \tag{P.11}$$

For a circuit in which any number of capacitors, n, are connected in series C_{equiv} is given by

$$C_{\text{equiv}} = \frac{1}{C_1} + \frac{1}{C_2} + \cdots + \frac{1}{C_n} \tag{P.12}$$

P.2.4 Resistors in Series and Parallel

Equivalent resistances in a circuit can be readily calculated through consideration of Kirchoff's laws. For resistors in series the equivalent resistance will be

$$R_{\text{equiv}} = R_1 + R_2 + R_3 + \cdots + R_n \tag{P.13}$$

Alternatively when resistance elements are arranged in parallel,

$$R_{\text{equiv}} = \frac{1}{R_1} + \frac{1}{R_2} + \cdots + \frac{1}{R_n} \tag{P.14}$$

These rules are used frequently to reduce a complicated arrangement of circuit elements to a much simpler equivalent circuit. Also they are essential in the study of charge transfer and the electrical properties of biological membranes.

P.2.5 RC Circuits and Relations

An important class of circuit is one in which both resistance and capacitance elements are found. These circuits exhibit time-varying behavior and are crucial systematic elements for the design of control circuits throughout nature. The behaviors of charge transfer in metabolism and in the nervous system show extensive use of RC circuit motifs (Fig. P.3).

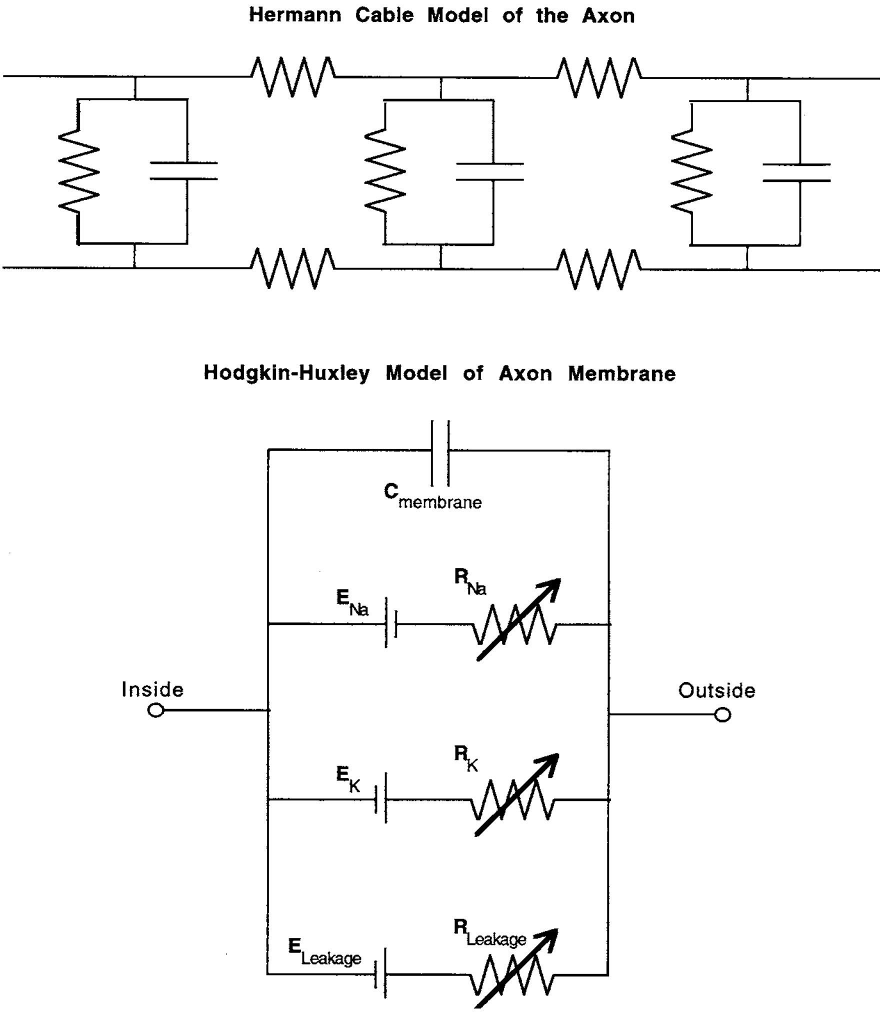

Hermann Cable Model of the Axon
Hodgkin-Huxley Model of Axon Membrane
C membrane
E Na
R Na
Inside
Outside
E K
R K
E Leakage
R Leakage

Fig. P.3 Models of the nerve cell membrane attempting to account for the measured voltage and current changes with the nerve impulse. (**a**) The Hermann cable model of the axon and (**b**) Hodgkin–Huxley Model of the axon membrane

Consider the circuit shown in Fig. P.4. The switch is placed in position A and current flows from the battery through the resistor and into the capacitor. Current will flow until adequate charge is separated on the plates of the capacitor to counter exactly the emf of the battery minus the potential dropped across the resistor. When the capacitor is fully charged, no current will flow in the circuit until the switch is placed in position B and the charge accumulated on the capacitor discharges through the resistor. The current then moves through the resistor in a reverse pattern. The time response of the charge and current flow with charging and discharging are

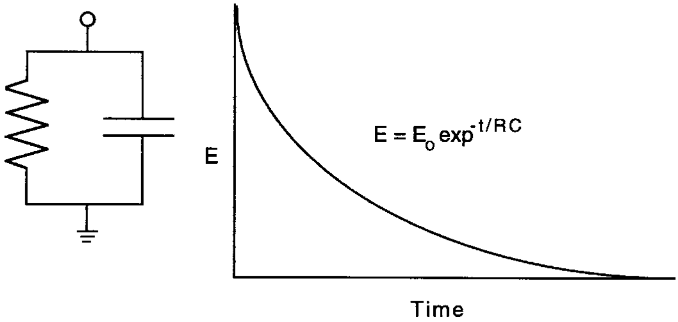


Fig. P.4 The voltage or current versus time relationships in an RC circuit. These relationships are fundamental to the behavior of cell membranes, charge transfer reactions, and information transfer in the nervous system

shown in Fig. P.4. The current flowing in the circuit at any time is given by

$$I = \frac{dQ}{dt} = \frac{E}{R} e^{-t} / RC \tag{P.15}$$

The current flow depends on the product *RC* which is called the *time constant*. Initially, at $t = 0$, a maximum current flows determined by E/R. Since there is no charge on the capacitor it offers no resistance to the initial flow of current. As charge builds on the capacitor, it develops a counter-emf and appears as an active resistance element in the circuit, which leads to an exponential fall in current flow with respect to time. When the switch is moved to position B, the capacitor discharges back through the resistor. The resistor serves as a limiting valve like the neck of a balloon that is blown up and then allowed to deflate. The time constant determines how fast the capacitor can be charged or discharged and, as Fig. P.4 shows at $t = 1$ RC, the voltage or current has fallen to $e^{-1} \approx 0.37$ times the initial current flow. Concurrently, the capacitor is $\approx$ 63% charged. By $3RC$ the capacitor is 95% charged and so on. The time relationships for discharging are equivalent but the current flow is reversed.

The *RC* circuit is a fundamental real-world element since all electrical systems contain some resistance and capacitance. Even a simple switch and resistor circuit have some element of capacitance and will have some time dependence as the switch is opened and closed. Similarly a capacitor has some intrinsic resistance, and its charging time must take these properties into account. For example, the charging time necessary for a copper conductor to reach electrostatic equilibrium will be on the order of 10^{-19} s which appears to be almost instantaneous. This extremely short time is due to the very low resistivity of the conductor, and the decay time for approaching electrostatic equilibrium will increase dramatically in systems using less conductive materials (such as will be found in biological charge transfer at membranes and in the semi-conducting circuits of computers).

P.3 Measuring Instruments

Electrical circuit measurements are familiar to all readers. Generally it is relatively simple to obtain a measuring device that can measure the potential difference (voltmeter), current flowing (ammeter), or resistance (ohmmeter) of a circuit or circuit elements. All of these instruments share an important practical and philosophical/epistemological problem: In order to measure some parameter of an electrical circuit or system, the measuring device is connected to the circuit. Thus it becomes a part of the circuit coupling the observer and the observed. This relationship can lead to apparent surprise and error. This relationship can be minimized if the interactions are made weak enough so that the observed system appears to be independent of the observer. The practical use of measuring instruments in electrical circuit analysis is governed by this principle (Fig. P.5). With this in mind, we can examine the mechanisms by which these measuring devices work and how the interactions between observer and observed are balanced depending on the observable under study.

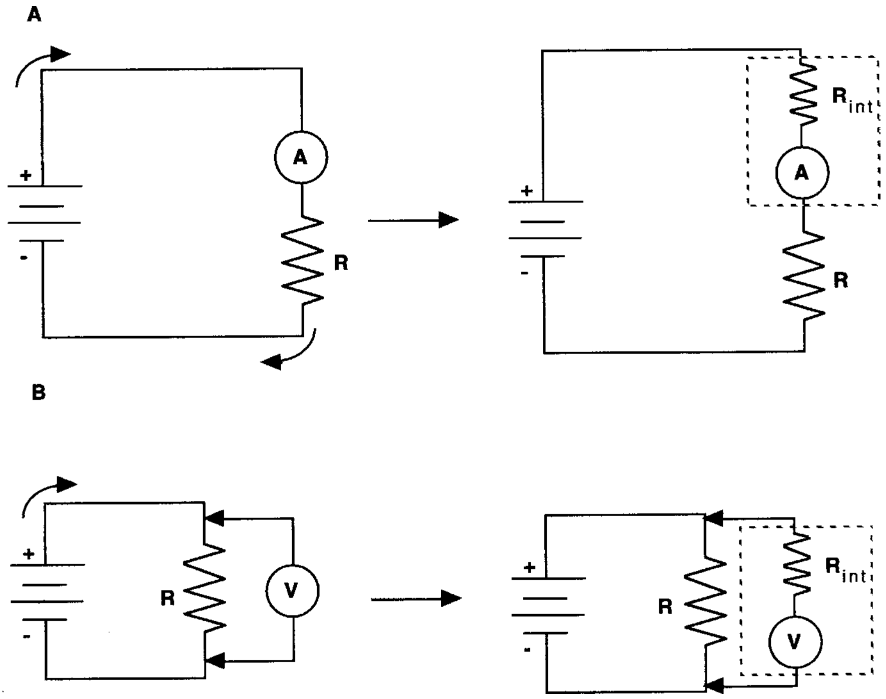


Fig. P.5 In (**a**) the arrangement for measuring the current flowing through a circuit loop with the equivalent internal resistance of the ammeter is shown; in (**b**) the arrangement of measuring the voltage difference across a circuit element with the internal resistance is shown

P.3.1 Ammeters, Voltmeters, Ohmmeters

Electrical measuring devices must have some means of indicating the value of the electric force, current, or resistance under study. In the cases of instruments with meters, the indicator is a *galvanometer* which is essentially a coil of wire suspended on a low friction movement in the field of a permanent magnet. When electric current flows in the coil a magnetic field is produced in the coil. The magnetized coil interacts with the permanent magnet inducing movement of the coil and its connected indicating needle. The degree of movement is proportional to the current flow through the coil. Modern digital instruments utilize transistorized circuitry to measure the current flow through the test rig.

An *ammeter* measures the current flow through a circuit wire and must be connected into the circuit in series in order to make the measurement (Fig. P.5a). If the ammeter has an internal resistance that is high it will load the circuit significantly and change the current flow in the circuit. In this case the ammeter will strongly interact with the circuit and will increase the likelihood of error and surprise being found in the observed system. The smaller the resistance of the ammeter, the more weakly it will interact with the natural system. The current flow in the circuit branch before inserting the ammeter is $I = E/R$, and the current in the branch after inserting the ammeter will be determined by the equivalent resistance of the ammeter in series with the circuit, $I = E/R{+}R_a$. Only when $R_a << R$ will the current in the circuit branch be unchanged by the insertion of the ammeter. Making current measurements can be practically challenging since a circuit must be interrupted and then reconnected with the ammeter in the circuit for the measurement to be possible.

A *voltmeter* measures the potential difference across a circuit element and to do so must be connected in parallel across the element under examination (Fig. P.5b). Since the voltmeter will only measure potential differences between two points, the circuit element under investigation must have resistance. If the voltmeter is to interact as weakly as possible with the natural system, it must have a high resistance so that minimal current is diverted from the circuit. This equivalent resistance is $\frac{1}{R_{\mathrm{equiv}}} = \frac{1}{R} + \frac{1}{R_{\mathrm{v}}}$. When $R_{\mathrm{v}} >> R$, then $R_{\mathrm{equiv}} \approx R$, and few elements of surprise and error will be introduced in the observable state space.

Appendix Q
Fermi's Golden Rule

A system is of a large number of microstates. When energy is transferred between a group of microstates in a system the overall energy will be exactly conserved but there will be some blurring in the arrangement of the approximate energy states. The degree of blurring depends on the nature of the perturbation. In general perturbations drive a system into a mixture of states. We cannot know the exact state B that an intial system A is driven to. This limit is related to Heisenberg's energy of uncertainty (10^{-27} J). Thus a transition from state i into a number of states in B is allowed and the rate at which the probability is transferred is calculated from an equation known as Fermi's Golden Rule. This rule determines the rate of the o*ne-to-many-jump*, P_{BA}. The rule has the following form:

$$P_{\mathrm{BA}} = \frac{2\pi}{\hbar} \left| H_{\mathrm{BA}} \right| g_{\mathrm{B}}$$

$|H_{\mathrm{BA}}|$ is the matrix element of the transition and is a complex number that characterizes the details of the perturbation. It has the dimensions of energy. g_{B} is the density of states in B. Thus this rule relates the kinetics of a quantum transition that depends on the form of the perturbation and the density of states and Planck's constant.

Appendix R
The Transition from Reactant to Product: Adiabatic and Non-adiabatic Transitions

In our earlier discussions of chemical transition state theory, it was assumed that once the transition state was achieved there would be almost spontaneous decay into product. From a classical thermodynamic standpoint this seems obvious but our concern is with the kinetic likelihood of the event occurring. In quantum mechanical systems, the likelihood of a transition is less certain. The movement of a system toward equilibrium is an irreversible process and occurs when a perturbation is applied to a system. Perturbations are time-dependent phenomena. The perturbing force, such as an electrical field may be applied slowly (i.e., at such a rate that the molecule is strongly coupled to the perturbation and thus responds completely). With this type of perturbation the molecule is induced to enter a new energy state with a very high rate of successful transition. In the language of quantum kinetics this is an *adiabatic transition*. The high probability of movement into the product state can be understood by considering the slow application of the perturbation force to slowly alter the wavefunction of the system from the reactant to the product state. If the process is slow enough the wavefunction adapts to the new parameters and the system moves to a new energy but remains in a single definite state. Thus there is no scattering into any other state except the new product state.

Alternatively if the perturbation is rapidly applied and removed, the system changes energy but the original wavefunction does not have time to change and the system retains the original wavefunction. This original wavefunction can be treated as the superposition of the wavefunctions of the new state and thus the interpretation is that the final state is a mixture of the wavefunctions of the new system. This situation arises during a rapid and impulsive transition and is called *non-adiabatic*. The consequence of the mixed wavefunctions is that in a non-adiabatic perturbation, the state of a system may move toward the transition state at B or it may approach the lower level of the excited energy surface such as C as shown in Fig. R.1. With a more rapid perturbation the system will be splayed and is more likely to miss the intersection of the reactant curve and the product curve (at B) and pass into an excited version of the reactant. This energy can be lost with a return to A rather than C with the result that in a non-adiabatic "reaction" there is a definite probability that no transition takes place. Thus in a non-adiabatic transition many attempts are made before a transition actually occurs.

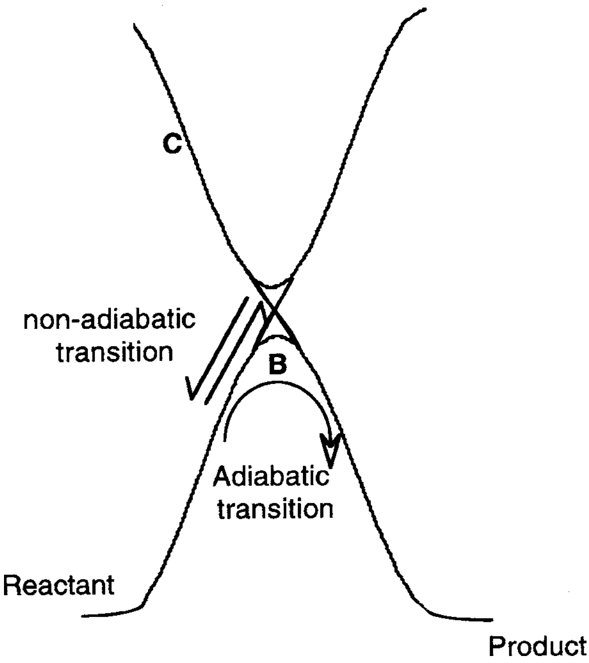


Fig. R.1 Perturbations may lead to adiabatic or non-adiabatic transitions

Index