"An excellent book for beginners and occasional practitioners"
Reviewer, Journal of the American Medical Association

Bioinformatics FOR DUMMIES®

2nd Edition

Updated to cover multiple new genomes and databases

A Reference for the Rest of Us!®

FREE eTips at dummies.com®

Jean-Michel Claverie, PhD
Research Director, France's Centre National de la Recherche Scientifique (CNRS)

Cedric Notredame, PhD
Professor of Bioinformatics, Switzerland's Lausanne University and the CNRS

Bioinformatics FOR DUMMIES®

2ND EDITION

by Jean-Michel Claverie, PhD
and Cedric Notredame, PhD

BICENTENNIAL 1807 WILEY 2007

Wiley Publishing, Inc.

Bioinformatics FOR DUMMIES®

2ND EDITION

Bioinformatics FOR DUMMIES®

2ND EDITION

by Jean-Michel Claverie, PhD
and Cedric Notredame, PhD

BICENTENNIAL 1807 WILEY 2007 BICENTENNIAL

Wiley Publishing, Inc.

Bioinformatics For Dummies®, 2nd Edition

Published by
Wiley Publishing, Inc.
111 River Street
Hoboken, NJ 07030-5774

www.wiley.com

Published by Wiley Publishing, Inc., Indianapolis, Indiana

Published simultaneously in Canada

LIMIT OF LIABILITY/DISCLAIMER OF WARRANTY: THE PUBLISHER AND THE AUTHOR MAKE NO REPRESENTATIONS OR WARRANTIES WITH RESPECT TO THE ACCURACY OR COMPLETENESS OF THE CONTENTS OF THIS WORK AND SPECIFICALLY DISCLAIM ALL WARRANTIES, INCLUDING WITHOUT LIMITATION WARRANTIES OF FITNESS FOR A PARTICULAR PURPOSE. NO WARRANTY MAY BE CREATED OR EXTENDED BY SALES OR PROMOTIONAL MATERIALS. THE ADVICE AND STRATEGIES CONTAINED HEREIN MAY NOT BE SUITABLE FOR EVERY SITUATION. THIS WORK IS SOLD WITH THE UNDERSTANDING THAT THE PUBLISHER IS NOT ENGAGED IN RENDERING LEGAL, ACCOUNTING, OR OTHER PROFESSIONAL SERVICES. IF PROFESSIONAL ASSISTANCE IS REQUIRED, THE SERVICES OF A COMPETENT PROFESSIONAL PERSON SHOULD BE SOUGHT. NEITHER THE PUBLISHER NOR THE AUTHOR SHALL BE LIABLE FOR DAMAGES ARISING HEREFROM. THE FACT THAT AN ORGANIZATION OR WEBSITE IS REFERRED TO IN THIS WORK AS A CITATION AND/OR A POTENTIAL SOURCE OF FURTHER INFORMATION DOES NOT MEAN THAT THE AUTHOR OR THE PUBLISHER ENDORSES THE INFORMATION THE ORGANIZATION OR WEBSITE MAY PROVIDE OR RECOMMENDATIONS IT MAY MAKE. FURTHER, READERS SHOULD BE AWARE THAT INTERNET WEBSITES LISTED IN THIS WORK MAY HAVE CHANGED OR DISAPPEARED BETWEEN WHEN THIS WORK WAS WRITTEN AND WHEN IT IS READ.

For general information on our other products and services, please contact our Customer Care Department within the U.S. at 800-762-2974, outside the U.S. at 317-572-3993, or fax 317-572-4002.

For technical support, please visit www.wiley.com/techsupport.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic books.

Library of Congress Control Number: 2006934844

ISBN13: 978-0-470-08985-9

ISBN10: 0-470-08985-7

Manufactured in the United States of America

10 9 8 7 6 5 4 3 2 1

1B/SX/RR/QW/IN

WILEY

About the Authors

Jean-Michel Claverie is Professor of Medical Bioinformatics at the School of Medicine of the Université de la Méditerranée, and a consultant in genomics and bioinformatics. He is the founder and current head of the Structural & Genomic Information Laboratory, located in Marseilles, a sunny city on the Mediterranean coast of France. Using science as a pretext to travel, Jean-Michel has held positions in Paris (France), Sherbrooke (PQ, Canada), the Salk Institute (La Jolla, CA), the Pasteur Institute (Paris), Incyte pharmaceutical (Palo Alto, CA); and the National Center for Biotechnology Information (Bethesda, MD). He has used computers in biology since the early days — his Ph.D. work involved modeling biochemical reactions by programming an 8K Honeywell 516 computer right from the console switches! Although he has no clear recollection of it, he has been credited with introducing the French word "bioinformatique" in the late eighties, before involuntarily coining the catchy "bioinformatics" by mistranslating it while giving a talk in English!

Jean-Michel's current research interests are in microbial and structural genomics, and in the development of bioinformatic methods for the prediction of gene function. He is the author or coauthor of more than 150 scientific publications, and a member of numerous international review panels and scientific councils. In his spare time, he enjoys the relaxed pace of life in Marseilles, with his wife Chantal and their two sons, Nicholas and Raphael.

Cedric Notredame is a researcher at the French National Centre for Scientific Research. Cedric has used and abused the facilities offered by science to wander around Europe. After a Ph.D. at EMBL (Heidelberg, Germany) and at the European Bioinformatics Institute (Cambridge, UK) under the supervision of Des Higgins (yes, the ClustalW guy), Cedric did a post-doc at the National Institute of Medical Research (London, UK), in the lab of Willie Taylor and under the supervision of Jaap Heringa. He then did a post-doc in Lausanne (Switzerland) with Phillip Bucher, and remained involved with the Swiss Institute of Bioinformatics for several years. Having had his share of rain, snow, and wind, Cedric has finally settled in Marseilles, where the sun and the sea are simply warmer than any other place he has lived in.

Cedric dedicates most of his research to the multiple sequence alignment problem and its many applications in biology. His friends claim that his entire life (past, present, future) is somehow stuffed into the T-Coffee multiple-sequence alignment package. When he is not busy dismantling T-Coffee and brewing new sequences, Cedric enjoys life in the company of his wife, Marita.

Dedication

This is for my parents Monique and Jack, for keeping me in school, and for Chantal, for keeping me happy — in and out of the lab. It's also for my daughter Vanessa, and my sons Nicholas and Raphael, for reminding me that not *everything* in life is scientific.

— J-MC

This is for my wife Marita, my daughter Lina, my mother Marie and in memory of my grandparents, Simone and Louis.

— CN

Authors' Acknowledgments

The entire Wiley staff did a great job pulling together to publish this book on tight deadlines. We'd especially like to thank our tireless project editor, Paul Levesque, and Barry Childs-Helton, who did a great job copyediting a text full of obscure biochemical words.

We'd also like to thank Amey Godse, our technical editor. Amey nailed down major and minor inaccuracies alike. His many suggestions did much to improve the book.

We also have to thank the bioinformatics community for creating the many great Web resources that we describe in this book and for making them available for free over the Internet. We personally know a number of the folks who keep these sites up and running — and salute all of them for their hard work, enthusiasm, and dedication. Topping this list are the staff members of the Swiss Bioinformatics Institute, who run the ExPASy and the Swiss EMBnet Web server. They always went out of their way to answer any query regarding their site. The NCBI folks have also been very helpful, and we thank them for that.

We also want to pat each other on the back for making the writing of this book great fun!

Finally, we'd like to thank our families and friends, who put up with missed dinners, extra child care, changing deadlines, late nights, and the many other demands of a project like this. We really appreciate their patience — and promise that we won't do another one . . . at least not anytime soon!

Publisher’s Acknowledgments

We’re proud of this book; please send us your comments through our online registration form located at www.dummies.com/register/.

Some of the people who helped bring this book to market include the following:

Acquisitions, Editorial, and Media Development

Project Editor: Paul Levesque

Acquisitions Editor: Melody Layne

Senior Copy Editor: Barry Childs-Helton

Technical Editor: Amey Godse

Editorial Manager: Leah Cameron

Media Development Specialists: Angela Denny, Kate Jenkins, Steven Kudirka, Kit Malone

Media Development Coordinator: Laura Atkinson

Media Project Supervisor: Laura Moss

Media Development Manager: Laura VanWinkle

Editorial Assistant: Amanda Foxworth

Sr. Editorial Assistant: Cherie Case

Cartoons: Rich Tennant (www.the5thwave.com)

Composition Services

Project Coordinator: Jennifer Theriot

Layout and Graphics: Carl Byers, Lavonne Cook, Barbara Moore, Shelley Norris, Barry Offringa, Laura Pence

Proofreaders: Susan Moritz, Charles Spencer, Rob Springer, Techbooks

Indexer: Techbooks

Anniversary Logo Design: Richard Pacifico

Publishing and Editorial for Technology Dummies

Richard Swadley, Vice President and Executive Group Publisher

Andy Cummings, Vice President and Publisher

Mary Bednarek, Executive Acquisitions Director

Mary C. Corder, Editorial Director

Publishing for Consumer Dummies

Diane Graves Steele, Vice President and Publisher

Joyce Pepple, Acquisitions Director

Composition Services

Gerry Fahey, Vice President of Production Services

Debbie Stailey, Director of Composition Services

Contents at a Glance

Table of Contents

Introduction

Welcome to the second edition of *Bioinformatics For Dummies!*

In the first edition, we presented bioinformatics as a brand new discipline on the rise. How right we were! Since then, it has become so prominent that anybody with an interest in biology, biotechnology, modern medicine, or (for that matter) genetically engineered food or drugs simply cannot afford to remain ignorant about the topic. With this book, you've come to the right place to quickly learn the basics.

But wait — if you expect something complicated, you're in for a (good or bad) surprise: Bioinformatics is nothing but good, sound, regular biology, appropriately dressed so it can fit into a computer.

Bioinformatics is about searching biological databases, comparing sequences, looking at protein structures, and (more generally) asking biological and biomedical questions with a computer. The bioinformatics we show you in this book can save you months of work in the lab at the minute cost of a few hours' work with your computer.

Although you'll find standard biological terms throughout, don't look here for long equations and computer-geek gibberish. The purpose of this book is to show you quickly and plainly how to use the bioinformatics programs that you need to get your work done. On every page, we give you tricks and treats to get the most out of existing tools. If you didn't know that you can use the most sophisticated programs for free over the Internet — and that you can do this (sometimes) without installing anything on your own computer — then stay tuned: You're in for many more good surprises.

What This Book Does for You

This book is here to help you get things done. For every standard bioinformatics task you may want to undertake, you'll find detailed steps that you can use to quickly produce the result you need.

To use most of the tools we describe in this book, you don't need to install any program on your computer. Everything we show you here runs over the Internet via your Internet browser.

If you know what you want to do — or at least know the task by name — going through the Table of Contents is the best strategy for finding exactly what you need. If you have an idea of what you want to do but you're not sure how to express it with words, Chapter 2 is here to help you decide which part of the book will suit your needs.

At the end of most chapters you'll find a convenient "Doing It for Free over the Internet" section, where we list a few carefully chosen Web sites that are similar to those we describe in the rest of the chapter. Treat this information as a spare wheel! If the main site is down, this section probably lists a convenient replacement.

Foolish Assumptions

Putting a project's assumptions right up front is just good policy. While writing this book, we have assumed that

- You have a PC running Microsoft Windows.
- You have an Internet connection (a fast one if possible, but not necessarily).
- You likely have a background in molecular biology. If you don't — or if you need to brush up on your molecular biology — Chapter 1 gives you a brief overview of the basics.
- You know how to use an Internet browser but not much more about computers.
- You don't want to become a bioinformatics guru; you simply want to use the right tools for your problem and not spend days finding out about things you don't need!

WARNING!

- Most private biotech companies consider it unsafe to send data over the Internet. We assume here that the data you want to analyze over the Internet is *not* very confidential. Also, some of the "public" databases and services listed in this book require commercial users to enter into a license agreement.

How This Book Is Organized

Bioinformatics is a broad field, with many nooks and crannies, hills and dales, and other charming features. Rather than present the whole vast discipline in one fell swoop, we've divided our discussion into five (more manageable) parts.

Part I: Getting Started in Bioinformatics

If you have less than an hour to find out what bioinformatics can do for you, Part I is the right place for you! It tells you everything you need to know in order to actually *do* something with bioinformatics. In Part I, we also remind you of just those bits of molecular biology that you'll need to know when you do sequence analysis. We show you here how to run the main bioinformatics tools so that you know what's in store for you.

Part II: A Survival Guide to Bioinformatics

If you want to find out everything that's ever been published on your sequence, this part is for you. It shows you how you can deal with the bioinformaticist's bread and butter: *DNA or protein sequences and their databases.* Here we tell you where you can find all the available sequences, and how to find the one you really need among zillions of irrelevant others. We also show you how to gather everything that's known in the universe about this special sequence that interests you so much (at least all of it that's available online).

Part III: Becoming a Pro in Sequence Analysis

If you want to compare sequences, this is the part for you. Here we show you how to search databases for sequences that are similar to yours, as well as show you how to compare two or more sequences. This part also tells you how to gather hints about the function of a gene, through sequence comparisons. Finally, we give you pointers on how to produce, edit, and beautify your multiple sequence alignments so you can show them in presentations and publications.

Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques

To take full advantage of this part, you should have a pretty good idea of what you're looking for. Heavy stuff is going on here: how to predict a protein structure, how to predict an RNA structure, and how to do phylogenetic analysis. These are complicated subjects; it's simply amazing what you can do with a simple PC, thanks to the Internet resources we describe in this part.

Part V: The Part of Tens

Welcome to our bazaar! If you haven't found what you were looking for in the other parts, you're now in the right place. The wealth of online resources that exist in bioinformatics is extraordinary — and almost overwhelming. With every student and his or her cousins putting semester reports online, finding exactly what you need with a simple keyword search can be a daunting task. In the Part of Tens, we give you a list of central resources that you can use as a starting point. Chances are that the program or server you're looking for is only one or two clicks away. In this part, we also give you ten important pieces of advice to make sure that your lab work can safely depend on your Internet work.

Icons Used in This Book

Always eager to please, we've decided to use a series of icons in the margins of this book as a way to help you key in on important information. We came up with four, which seemed like a nice, round number.

TECHNICAL STUFF

Some particularly technoid information is coming up. You can skip it and nothing terrible will happen. Yet, if you want to be in full control of what you're doing, reading this may help! Your call. . . .

TIP

This icon shows you something simple, or smart, or a cute shortcut. In any case, it's something that can save you time and trouble.

WARNING!

There are many booby traps around when you use Internet servers. This icon warns you when some ambiguity surrounds what the server you're using is up to — or when disaster is only one (wrong) mouse click away. Treat the Warning icon with respect — especially in a steps list!

REMEMBER

This icon indicates something you should remember. It can be one of the few important principles that you need to know, or it can be a very special tip — the kind that can save you three days of work (or drive you nuts if you forget it). You may assume that the head of your institute/company got to the top by discovering and applying one or more of pearls of wisdom in these very special tips!

Where to Go from Here

If you know nothing about bioinformatics, this book is here to reassure you. Bioinformatics is a much simpler subject than you ever thought possible. For most people new to this field, the main difficulty is finding out the kind of questions they can ask with these new tools. If you're a biologist, don't let the computer scare you; bioinformatics is nothing more than good, sound, regular biology hidden inside a computer.

The magic thing about bioinformatics is that, with a simple Internet connection, you can browse databases that contain the sum of our entire human biological knowledge — and you can do this with the most sophisticated tools ever developed by mankind. And how much is this going to cost you? Nothing!

If you do molecular biology, this is the equivalent of having an entire lab with expensive, state-of-the-art equipment and staffed by an army of post-docs who can go fetch anything you need any time you need it. The only difference is that you cannot set this lab on fire (even if you try very hard).

If you think of it, it is quite incredible to realize that all this is right here, at your fingertips, one or two mouse clicks away! The Web is borderless; it is colorblind and unimpressed by wealth! Whether you come from a rich or a poor country, whether you're a first-year student, a scientist, or a Nobel Prize winner, you have access — for free — to the same high-quality information. No other scientific discipline has ever been so democratically widespread.

This book isn't a textbook but a cookbook! And we take pride in this! It contains many recipes that colleagues showed us over the years or that we discovered ourselves. Accommodating and serving biological data is something very personal — and we're sure that you'll gradually find your own way to do it. In the meantime, if you need a quick fix, you can always use some of the off-the-shelf solutions that we provide here.

No discipline in science has benefited as much as biology from the "global village" phenomenon of the Internet. Whatever your question, whatever you want to do, starting on the Internet is the proper thing to do. Nonetheless, remember that the best *and* the worst appear online these days. Do as you do in real life — and trust only those sites or institutions that you know well.

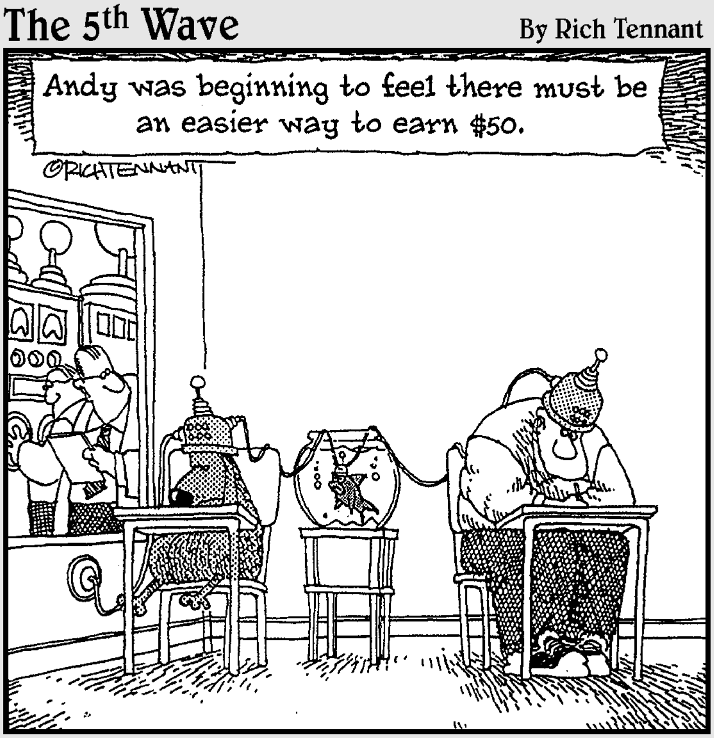
TIP

This book is as up-to-date as we can make it, but the world doesn't stand still right after we finish correcting the last galley proofs and send *Bioinformatics For Dummies* into the bookstores. For those of you who want up-to-date info on the growing field of bioinformatics (including lists of our favorite bioinformatics links) and don't want to wait until the next edition, check out the Web site associated with this title at `www.dummies.com/extras`.

Sometimes browsing the Internet gives one the depressing feeling that everything has been done by others and that it's all over. This may be true. Now that the whole world talks together, it's clear that there's a finite number of interesting questions to ask. That's the bad news. The good news is that there are many more answers than there are questions! Never exclude the hypothesis that your answer may be the best in the universe (at least for a few days. . . .)!

Part I

Getting Started in Bioinformatics


The 5th Wave
By Rich Tennant
Andy was beginning to feel there must be an easier way to earn $50.
©RICHTENNANT

In this part . . .

Bioinformatics is a new discipline, which means that nobody should feel ashamed if he or she doesn't have a clue what the excitement's all about. Don't worry; after finishing this book, you'll be speaking bioinformaticsspeak with the best of them.

We start you off in Part I with a quick reminder of what you need to know about DNA and proteins to make sense of this book. We also give you an overview of the main bioinformatics tools available on the Internet.

We don't give too many details here, but if all you need to know is which Internet page to open and which button to press, come on in, 'cuz we've got just what you need!

Chapter 1

Finding Out What Bioinformatics Can Do for You

In This Chapter

- Defining bioinformatics
- Understanding the links between modern biology, genomics, and bioinformatics
- Determining which biological questions bioinformatics can help you answer quickly

> *Organic chemistry is the chemistry of carbon compounds. Biochemistry is the study of carbon compounds that crawl.*
>
> — Mike Adam

It looks like *bio*logists are colonizing the dictionary with all these *bio*-words: we have bio-chemistry, bio-metrics, bio-physics, bio-technology, bio-hazards, and even bio-terrorism. Now what's up with the new entry in the bio-sweepstakes, bio-informatics?

What Is Bioinformatics?

In today's world, computers are as likely to be used by biologists as by any other highly trained professionals — bankers or flight controllers, for example. Many of the tasks performed by such professionals are common to most of us: We all tend to write lots of memos and send lots of e-mails; many of us use spreadsheets, and we all store immense amounts of never-to-be-seen-again data in complicated file systems.

However, besides these general tasks, biologists also use computers to address problems that are very specific to biologists, which are of no interest to bankers or flight controllers. These specialized tasks, taken together, make up the field of *bioinformatics.* More specifically, we can define bioinformatics as the computational branch of molecular biology.

Time for a little bit of history. Before the era of bioinformatics, only two ways of performing biological experiments were available: within a living organism (so-called *in vivo*) or in an artificial environment (so-called *in vitro,* from the Latin *in glass*). Taking the analogy further, we can say that bioinformatics is in fact *in silico* biology, from the silicon chips on which microprocessors are built.

This new way of doing biology has certainly become very trendy, but don't think that "trendy" translates into "lightweight" or "flash-in-the-pan." Bioinformatics goes way beyond trendy — it's at the center of the most recent developments in biology, such as the deciphering of the human genome (another buzzword), "system biology" (trying to look at the global picture), new biotechnologies, new legal and forensic techniques, as well as the personalized medicine of the future.

Because of the centrality of bioinformatics to cutting-edge developments in molecular biology, people from many different fields have been stumbling across the term in a variety of different contexts. If you're a biology, medical, or computer science student, a professional in the pharmaceutical industry, a lawyer or a policeman worrying about DNA testing, a consumer concerned about GMOs (Genetically Modified Organisms), or even a NASDAQ investor interested in start-up companies, you'll already have come across the word *bioinformatics.* If you're good at what you do, you'll want to know what all the fuss is about. This chapter, then, is for you.

Instead of a formal definition that would take hours to cover all the ins and outs of the topic, the best way to get a quick feel for what bioinformatics — or swimming, for that matter — is all about is to jump right into the water; that's what we do next. Go ahead and get your feet wet with some basic molecular biology concepts — and the relevant questions intimately connected with such concepts — that all together define bioinformatics.

Analyzing Protein Sequences

If you eat steak, you're intimately acquainted with proteins. (Your taste buds know them intimately anyway, even if your rational mind was too busy with dinner to master the concept.) For you non-steak lovers out there, you'll be pleased to know that proteins abound in fish and vegetables, too. Moreover, all these proteins are made up of the same basic building blocks, called *amino acids.* Amino acids are already quite complex organic molecules, made of carbon, hydrogen, oxygen, nitrogen, and sulfur atoms. So the overall recipe for a protein (the one your rational mind will appreciate, even if your taste buds won't) is something like $C_{1200}H_{2400}O_{600}N_{300}S_{100}$.

The early days of biochemistry were devoted to finding out a better way to represent proteins — preferably in terms of a formula that would explain their biological (or even nutritional) properties. Biochemists realized over time that proteins were huge molecules (*macromolecules*) made up of large numbers of amino acids (typically from 100 to 500), picked out from a selection of 20 "flavors" with names such as alanine, glycine, tyrosine, glutamine, and so on. Table 1-1 gives you the list of these 20 building blocks, with their full names, three-letter codes, and one-letter codes (the *IUPAC code,* after the *International Union of Pure and Applied Chemistry* committee that designed it).

Table 1-1 The 20 Amino Acids and Their Official Codes

#	*1-Letter Code*	*3-Letter Code*	*Name*
1	A	Ala	Alanine
2	R	Arg	Arginine
3	N	Asn	Asparagine
4	D	Asp	Aspartic acid
5	C	Cys	Cysteine
6	Q	Gln	Glutamine
7	E	Glu	Glutamic acid
8	G	Gly	Glycine
9	H	His	Histidine
10	I	Ile	Isoleucine
11	L	Leu	Leucine
12	K	Lys	Lysine
13	M	Met	Methionine
14	F	Phe	Phenylalanine
15	P	Pro	Proline
16	S	Ser	Serine
17	T	Thr	Threonine
18	W	Trp	Tryptophan
19	Y	Tyr	Tyrosine
20	V	Val	Valine

Biochemists then recognized that a given type of protein (such as insulin or myoglobin) always contains precisely the same number of total amino acids (generically called *residues*) — in the same proportion. Thus, a better formula for a protein looks like this:

insulin = (30 glycines + 44 alanines + 5 tyrosines + 14 glutamines + . . .)

Finally, biochemists discovered that these amino acids are linked together as a chain — and that the true identity of a protein is derived not only from its composition, but also from the precise order of its constituent amino acids. The first amino-acid sequence of a protein — insulin — was determined in 1951. The actual recipe for human insulin, from which all its biological properties derive, is the following chain of 110 residues:

insulin = MALWMRLLPLLALLALWGPDPAAAFVNQHLCGSHLVEALYLVCGERG
FFYTPKTRREAEDLQVGQVELGGGPGAGSLQPLALEGSLQKRGIVEQCCTSICSLY
QLENYCN

Now, more than 50 years later, analyzing protein sequences like these remains a central topic of bioinformatics in all laboratories throughout the world. (Check out Chapters 2, 4, and 6 through 11 to quickly figure out how to analyze your protein sequence and become a member of the club!)

A brief history of sequence analysis

Besides earning Alfred Sanger his first Nobel Prize, the sequencing of insulin inaugurated the modern era of molecular and structural biology. Traditionally a *soft science* (that is, more tolerant of fuzzy reasoning and hand-waving ambiguity than chemistry or physics), biology got a taste of its first fundamental dataset: molecular sequences. In the early 1960s, known protein sequences accumulated slowly — perhaps a blessing in disguise, given that the computers capable of analyzing them hadn't been developed! In this pre-computer era (from our present perspective, anyway), sequences were assembled, analyzed, and compared by (manually) writing them on pieces of paper, taping them side by side on laboratory walls, and/or moving them around for optimal alignment (now called *pattern matching*).

As soon as the early computers became available (as big as locomotives and just as fast, and with 8K of RAM!), the first computational biologists started to enter these manual algorithms into the memory banks. This practice was brand new — nobody before them had to manipulate and analyze molecular sequences as *texts*. Most methods had to be invented from scratch, and in the process, a new area of research — the analysis of protein sequences using computers — was generated. This was the genesis of bioinformatics.

Seven additional amino acid codes

When you work with databases or analysis programs, you're likely to have some unusual letters popping up now and then in your protein sequences. These letters are either used to designate exotic amino acids, or are used to denote various levels of ambiguity — that is, a total lack of information — about certain positions in the sequence. We've listed these particular letters in the following table.

Seven Codes for Ambiguity or Exceptional Amino Acids

1-Letter Code	*3-Letter Code*	*Meaning*
B	Asn or Asp	Asparagine or aspartic acid
J	Xle	Isoleucine or leucine
O (letter)	Pyl	Pyrrolysine
U	Sec	Selenocysteine
Z	Gln or Glu	Glutamine or glutamic acid
X	Xaa	Any residue
--	-----	No corresponding residue (gap)

The B and Z codes (which are now becoming obsolete) indicated how hard it was to distinguish between Asp and Asn (or Glu and Gln) in the early days of protein sequence determination. In contrast, the J code shows how difficult it is to distinguish between Ile and Leu using mass spectrometry, the latest sequencing technique. The Pyl and Sec exotic amino acids are specified by the UAG (Pyl) and UGA (Sec) stop codons read in a specific context. The *X* code is still very much used as a placeholder letter when you don't know the amino acid at a given position in the sequence. Alignment programs use "–" to denote positions apparently missing from the sequence.

Reading protein sequences from N to C

The twenty amino-acid molecules found in proteins have different *bodies* (their characteristic residues, listed in Table 1-1) — but all have the same pair of *hooks* — NH_2 and COOH. These groups of atoms are used to form the so-called *peptidic bonds* between the successive residues in the sequence. Figure 1-1 shows free individual amino acids floating about, displaying their hooks for all to see.

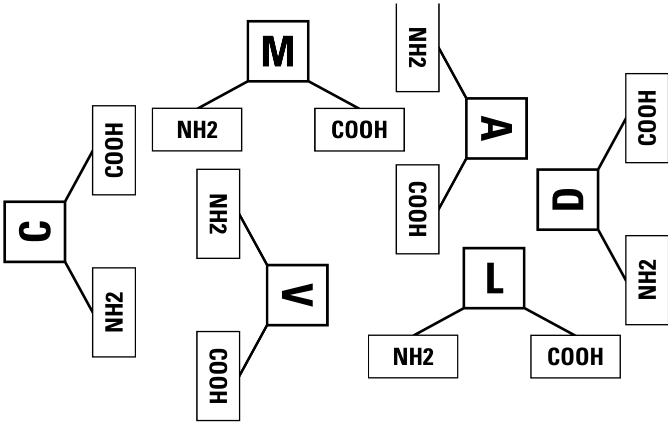


Figure 1-1: Free amino acids floating around.

The protein molecule itself is made when a free NH_2 group links chemically with a COOH group, forming the peptide bond CO-NH. Figure 1-2 shows a schematic picture of the resulting chain.

As a result of this chaining process, your protein molecule is going to be left with an unused NH_2 at one end and an unused COOH at the other end. These extremities are called (respectively) the *N-terminus* and *C-terminus* of the protein chain. This is important to know because scientific convention (in books, databases, and so on) defines the *sequence* of a protein — or of a protein fragment — as the succession of its constituent amino acids, listed in order from the N-terminus to the C-terminus. The sequence of our (short!) demo protein is then

```
MAVLD= Met-Ala-Val-Leu-Asp= Methionine-Alanine-
          Valine-Leucine-Aspartic
```

Working with protein 3-D structures

The precise succession of a protein's constituent amino acids is what defines a given protein molecule. This ribbon of amino acids, however, is not what

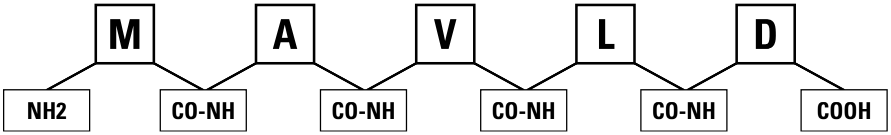


Figure 1-2: Amino acids chained together to constitute a protein molecule.

gives the protein its biological properties (for instance, its ability to digest sugar or to become part of a muscle fiber); those come from the three-dimensional (3-D) shape that the ribbon adopts in its environment. A protein molecule, once made, is not a chainlike, highly flexible object (think like a section of chain-link fence); rather, it's more like a compact, well-bundled ball of string. The final 3-D shape of the protein molecule is uniquely dictated by its sequence because some amino-acid types (for instance, *hydrophobic* residues L, V, I) have no desire whatsoever to be at the surface interacting with the surrounding water — while others (for instance, *hydrophilic* residues D, S, K) are actively looking for such an opportunity. The protein chain is also affected by other influences, such as the electric charges carried by some of the amino acids, or their capacity to fit with their immediate neighbors.

The first 3-D structure of a protein was determined in 1958 by Drs. Kendrew and Perutz, using the complicated technique of X-ray crystallography. (Not for the faint of heart. Don't grapple with how it works unless you want to turn professional!) Besides winning one more Nobel Prize for the nascent field of molecular biology, this feat made the doctors realize that proteins have precise and specific shapes, encoded in the sequence of amino acids. Hence, they predicted that proteins with similar sequences would fold into similar shapes — and, conversely, that proteins with similar structures would be encoded by similar sequences of amino acids. The function of a protein turned out to be a direct consequence of its 3-D structure (shape). The resulting logical linkage

```
SEQUENCE⇨STRUCTURE⇨FUNCTION
```

was established, and is now a central concept of molecular biology and bioinformatics.

Playing with protein structure models on a computer screen is, of course, much easier than carrying around a thousand-piece, 3-D plastic puzzle. As a consequence, an increasing proportion of the bioinformatics pie is now devoted to the development of cyber-tools to navigate between sequences and 3-D structures. (This specialized area is called *structural bioinformatics.*) Thanks to many free resources on the Internet, it is not difficult to display some beautiful protein pictures on your own computer — and start playing with them as in video games. (We show you how to do that in Chapter 11.)

Before you get a chance to read that chapter, Figure 1-3 gives you an idea of what a 400-amino-acid typical protein 3-D (schematic) structure looks like — when you don't have a color monitor and can't make it move and turn!

REMEMBER

Don't forget: Protein molecules, even in their wonderful complexity, are still pretty small. The one in Figure 1-3 would fit in a box whose sides measure 70/1,000,000 millimeters. There are thousands of different proteins in a single bacterium, each of them in thousands of copies — more than enough evidence that Life Is Not Simple!

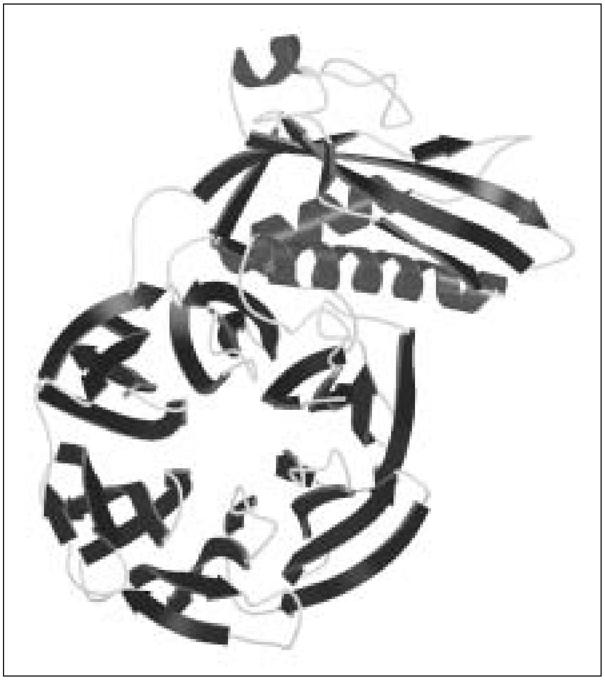


Figure 1-3: Example of protein 3-D structure (schematic).

Protein bioinformatics covered in this book

The study of protein sequences can get pretty complicated — so complicated, in fact, that it would take a pretty thick book to cover all aspects of the field. We'd like to take a more selective approach by focusing on those aspects of protein sequences where bioinformatic analyses can be most useful. The following list gives you a look at some topics where such an analysis is particularly relevant to protein sequences — and also tells which chapters of this book cover those topics in greater detail:

- Retrieving protein sequences from databases (Chapters 2, 3, and 4)
- Computing amino-acid composition, molecular weight, isoelectric point, and other parameters (Chapter 6)
- Computing how hydrophobic or hydrophilic a protein is, predicting antigenic sites, locating membrane-spanning segments (Chapter 6)
- Predicting elements of secondary structure (Chapters 6 and 11)
- Predicting the domain organization of proteins (Chapters 6, 7, 9, and 11)
- Visualizing protein structures in 3-D (Chapter 11)
- Predicting a protein's 3-D structure from its sequence (Chapter 11)

- Finding all proteins that share a similar sequence (Chapter 7)
- Classifying proteins into families (Chapters 7, 8, and 9)
- Finding the best alignment between two or more proteins (Chapters 8 and 9)
- Finding evolutionary relationships between proteins, drawing proteins' family trees (Chapters 7, 9, 11, and 13)

Analyzing DNA Sequences

During the 1950s, while scientists such as Kendrew and Perutz were still struggling to determine the first 3-D structures of proteins, other biologists had already acquired a lot of indirect evidence (via extremely clever genetics experiments) that *deoxyribonucleic acid* (DNA) — the stuff that makes up our genes — was *also* a large macromolecule. It was a long, chainlike molecule twisted into a double helix, and each link in the chain was a pairing of two out of four constituents called *nucleotides.* (A nucleotide is made up of one phosphate group linked to a pentose sugar, which is itself linked to one of 4 types of nitrogenous organic bases symbolized by the four letters A, C, G, and T.)

However, molecular biologists had to wait until much later — the 1970s, to be more precise — before they could determine the sequence of DNA molecules and get direct access to the sequences of gene nucleotides.

This was a revolution (earning A. Sanger his second Nobel Prize!) because the small DNA sequence alphabet (4 nucleotides, as compared to 20 amino acids) allowed a much simpler and faster reading — and quickly lent itself to complete automation. Currently, the worldwide rate of determining DNA sequences is faster (by orders of magnitude) than the rate of protein sequencing.

Reading DNA sequences the right way

As was the case for the 20 amino acids found in proteins, the 4 nucleotides making DNA have different bodies but all have the same pair of hooks: 5' phosphoryl and 3' hydroxyl (pronounced *five prime* and *three prime*) by reference to their positions in the deoxyribose sugar molecule, which is part of the nucleotide chaining device. Figure 1-4 shows what free individual nucleotides look like.

Forming a bond between the 5' and 3' positions of the constituent nucleotides then makes the DNA molecule. Figure 1-5 shows a schematic representation of the resulting DNA strand.

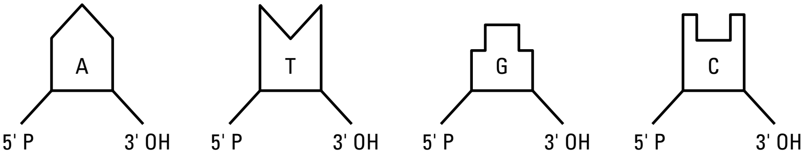


Figure 1-4: The four nucleotides making DNA.

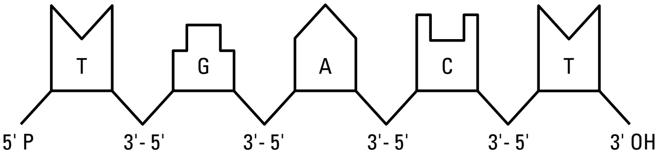


Figure 1-5: Chained nucleotides constituting a DNA strand.

After the nucleotides are linked, the resulting DNA strand exhibits an unused phosphoryl group (PO_4) at the 5' end, and an unused hydroxyl group (OH) at the 3' end. These extremities are respectively called the *5'-terminus* and the *3'-terminus* of the DNA strand.

A DNA sequence is always defined (in books, databases, articles, and programs) as the succession of its constituent nucleotides *listed from the 5'- to 3'- terminus* (that is, end). The sequence of the (short!) DNA strand shown in Figure 1-5 is then

```
TGACT = Thymine-Guanine-Adenine-Cytosine-Thymine
```

The two sides of a DNA sequence

In the same laboratory where Kendrew and Perutz were trying to figure out the first 3-D structure of a protein, Watson and Crick elucidated — in 1953 — the famous double-helical structure of the DNA molecule. These days everybody has a mental picture of this famous spiral-staircase molecule; the elegance of the DNA *double helix* probably helped make it the most popular notion to come out of molecular biology. But what made this discovery so important — earning one more Nobel Prize for molecular biology — was not the helical shape, but the discovery that the DNA molecule consists of two complementary strands, shown in Figure 1-6.

REMEMBER

The IUPAC code for DNA sequences

The following table lists the one-letter codes (IUPAC codes) used to work with DNA sequences. Official IUPAC codes, from the International Union of Pure and Applied Chemistry, are defined for all possible two- and three-way ambiguities. The table shows only the ones most frequently used.

Most Common Letters Used for DNA Nucleotide Sequences

1-Letter Code	*Nucleotide Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
T	Thymine	Pyrimidine
N	Any nucleotide (any base)	(n/a)
R	A or G	Purine
Y	C or T	Pyrimidine
--	-----	None (gap)

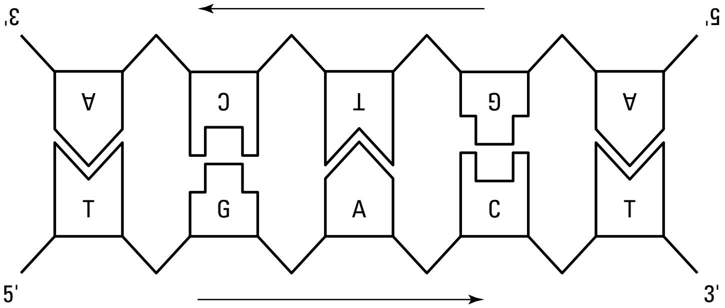


Figure 1-6: The two complementary strands of a complete DNA molecule.

By *complementarity,* we mean that a thymine (T) on one strand is always facing an adenine (A) (and vice versa) — and guanine (G) is always facing a cytosine (C). These couples, A-T and G-C, although not linked by a chemical bond, have a strict one-to-one reciprocal relationship. When you know the sequence of nucleotides along one strand, you can automatically deduce the sequence on the other one. This amazing property — and not the stylish helical structure — is the Rosetta Stone that explains everything about DNA sequences. For instance, when living organisms reproduce, each of their genes must be duplicated. In order to do this, nature doesn't go about it the way a photocopier would — by making an exact copy. Rather, nature separates the DNA strands and makes *two complementary* ones, thanks to the magical two-sided structure of DNA molecules.

This double strand structure of DNA makes the definition of a DNA sequence ambiguous: Even with our convention of reading the nucleotides from the 5' end toward the 3' end, you may decide to write down the bottom or the top sequence. Convince yourself that they're both equally valid sequences by turning this book upside down! Thus, at each location, a DNA molecule corresponds to two — totally different — sequences, related by this reverse-and-complement operation. This isn't complicated; simply keep it in mind every time you work with DNA sequences.

Fortunately, most database mining programs, such as BLAST, know about this property, and take both strands into account when reporting their results. But some programs don't bother — and only analyze the sequence you gave them. In cases where both strands matter, always make sure that a complete analysis has been performed. (We discuss these details further in Chapters 3, 5, and 7.)

Palindromes in DNA sequences

Newcomers to DNA sequence analysis are usually confused by the notion of reverse complementary sequences. However, in due time you'll be able to recognize right away that the two sequences

```
ATGCTGATCTTGGCCATCAATG and CATTGATGGCCAAGATCAGCAT
```

correspond to facing strands of the same DNA molecule.

One fascinating property of DNA complementarity is the fact that regions of DNA may correspond to sequences that are identical when read from the two complementary strands. Figure 1-7 helps illustrate this magic trick.

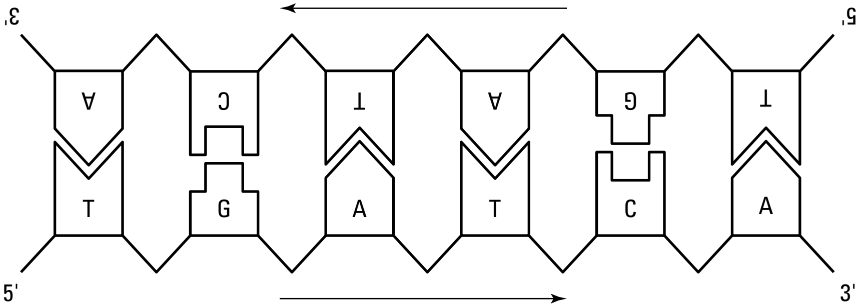


Figure 1-7: How two complementary strands can be read the same way.

Such sequences are called *palindromes,* after the term for a phrase or sentence that reads the same in both directions (such as “Madam, I’m Adam” or “A man, a plan, a canal: Panama.”) *Palindromic* sequences aren’t merely a curiosity; they play important biological roles. For instance, most DNA cutting enzymes (so-called *restriction enzymes*) have palindromic target sequences. Other palindromic sequences serve as *binding sites,* where regulatory proteins stick so they can turn genes on and off. Palindromic sequences also have a strong influence on the 3-D structure of DNA molecules. (And not just DNA. See the next section for more on palindromic sequences in RNA.) Looking for exact or approximate palindromes in DNA sequences is a classic bioinformatic exercise.

Analyzing RNA Sequences

DNA (deoxyribonucleic acid) is the most dignified member of the nucleic acid family of macromolecules. Its sole and only task is to ensure — forever — the conservation of the genetic information for its organism. It is thus very stable and resistant, and lies well-protected in the nucleus of each cell. *Ribonucleic acid* (RNA) is a much more active member of the nucleic acid family; it’s synthesized and degraded constantly as it makes copies of genes available to the cell factory.

In the context of bioinformatics, there are only two important differences between RNA and DNA:

- RNA differs from DNA by one nucleotide.
- RNA comes as a single strand, not a helix.

The one-letter IUPAC codes for RNA sequences are shown in Table 1-2.

Table 1-2 Most Common Letters Used for RNA Nucleotide Sequences

1-Letter Code	*Nucleotide Base Name*	*Category*
A	Adenine	Purine
C	Cytosine	Pyrimidine
G	Guanine	Purine
U	Uracil	Pyrimidine
N	Any nucleotide	Purine or Pyrimidine

(continued)

Table 1-2 *(continued)*

1-Letter Code	*Nucleotide Base Name*	*Category*
R	A or G	Purine
Y	C or U	Pyrimidine
--	-------	None (gap)

Some programs automatically handle the U-instead-of-T conversion — and many don't even distinguish between the two classes of nucleic acids. So don't be surprised if a database entry displays RNA sequences (such as messenger RNA) with a T instead of a U. In fact, like proteins, RNA sequences are encoded in the DNA. For this reason, people have adopted the habit of working with the sequences of the RNA *genes* (written in DNA) rather than with RNA sequences.

RNA structures: Playing with sticky strands

Even though RNA molecules consist of single strands of nucleotides, their natural urge for pairing with complementary sequences is still there. Think of each such single strand as a free-floating piece of Scotch tape: You know that it won't take long for that tape to become a messy ball, until no sticky part remains exposed. This is exactly what happens to the single-stranded RNA molecule — more or less (for the sake of poetic license) — although Figure 1-8 shows more precisely how the stickiness works.

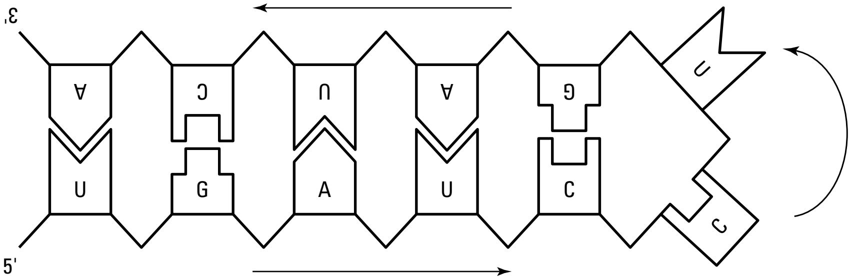


Figure 1-8: How RNA turns itself into a double-stranded structure.

Now you understand why we insisted on the notion of strand complementarity (refer to Figure 1-6). Single-stranded RNA molecules pair different regions of their sequences to form stable double-helical structures — admittedly less regular than (but quite similar to) the double-helical structure of DNA. Once synthesized, each RNA molecule quickly adopts a compact fold — trying to pair as many nucleotides as possible, while keeping the chain not only flexible but true to its own geometry. Hairpin shapes, as shown in Figure 1-8, are the basic elements of RNA secondary structure; they're made up of *loops* (the unpaired C-U in Figure 1-8) and *stems* (the paired regions).

Just for fun, verify for yourself that a palindromic RNA sequence results in a perfect hairpin, with no loop. While attempting to pair as many nucleotides as possible, the RNA chain folds in space, resulting in a specific 3-D structure that's dictated by its sequences. As with proteins, the linear sequence of the building blocks dictates the final 3-D shape. The biological function of RNA molecules derives from their 3-D shapes or from their sequence complementarity with specific genes.

Computing (predicting) the final fold of an RNA molecule from its sequence is a challenging problem that drove many historical developments in bioinformatics. The recent discovery that small RNA molecules can switch off the activity of a number of genes is what triggered a renewed interest in these sticky sequences. (Go directly to Chapter 12 if your main interest is in RNA bioinformatics.)

More on nucleic acid nomenclature

Don't panic if you get the impression that books, courses, and the technical literature all use many different words and abbreviations to designate the building blocks of nucleic acids: That's actually true — for example, you'll find "base," "base pair," "nucleoside," and "nucleotide" — but note: These different names designate slightly different chemical entities, and those differences are irrelevant for us just now. So far we've used the term *nucleotide* — abbreviated *nt* (as in "a 400-nt-long sequence"). This way of labeling a sequence refers to the length of the DNA (or RNA) molecules in terms of the number of positions they have available for nucleotides. For instance, the sequence in Figure 1-5 is 5 nt long.

Notice that we say *number of positions* rather than *number of nucleotides.* A 400-nt long DNA molecule has 400 positions for nucleotides, but it actually contains twice that many (800) because every position contains a pair of nucleotides. To make this clearer, DNA sequence sizes are often given in *base pairs,* abbreviated *bp.* Thus the DNA sequence in Figure 1-5 is 5 bp long. Larger units, such as *kb* (1000 bp) or *Mb* (mega-bp) are also used.

DNA Coding Regions: Pretending to Work with Protein Sequences

Of the hundreds of thousands of protein sequences found in current databases, only a small percentage correspond to molecules that have actually been isolated by somebody or experimented upon. That's because determining the sequence of a protein is much more difficult than sequencing DNA — but all the proteins that a given organism (whether microbe or human being) can synthesize are encoded in the DNA sequence of its genome. Thus, the smart shortcut that molecular biologists have been using is to read protein sequences directly at the information source: in the DNA sequence! This way, we can pretend to know the amino-acid sequence of a protein that has never been isolated in a test tube.

Turning DNA into proteins: The genetic code

When you know a DNA sequence, you can translate it into the corresponding protein sequence by using the genetic code, the very same way the cell itself generates a protein sequence. The genetic code is universal (with some exceptions — otherwise life would be too simple!), and it is nature's solution to the problem of how one uniquely relates a 4-nucleotide sequence (A, T, G, C) to a suite of 20 amino acids; we're using symbols (rather than actual chemicals) to do the same. Understanding how the cell does this was one of the most brilliant achievements of the biologists of the 1960s. Yet the final answer can be contained in a (miraculously small) table — as shown in Figure 1-9. Have a look, but feel free to indulge in awed silence as you enter the most sacred monument of modern biology.

Here's how to use the table shown in Figure 1-9: From a given starting point in your DNA sequence, start reading the sequence 3 nucleotides (one *triplet*) at a time. Then consult the genetic code table to read which amino acid corresponds to the current triplet (technically referred to as *codons*). For instance, the following DNA (or messenger RNA) sequence is decoded as follows:

1. **Read the DNA sequence:**

   ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG

2. **Decompose it into successive triplets:**

   ATG GAA GTA TTT AAA GCG CCA CCT ATT GGG ATA TAA G . . .

3. **Translate each triplet into the corresponding amino acid:**

   M E V F K A P P I G I STOP

If your DNA sequence is correctly listed in the 5' to 3' orientation, you generate the protein sequence in the conventional N- to C-terminus as well. This approach has an advantage: You don't have to think about these orientation details ever again.

Thus, if you know where a protein-coding region starts in a DNA sequence, your computer can pretend to be a cell and generate the corresponding amino-acid sequence! This simple computer translation exercise is at the origin of most of the so-called *protein sequences* that you can find in databases. Many sequence analysis programs acknowledge this fact by offering on-the-fly translation, so you can process DNA sequences as virtual protein sequences with a simple mouse click.

More with coding DNA sequences

Using the example in the first paragraphs of the section "DNA Coding Regions: Pretending to Work with Protein Sequences," you can see that the resulting protein sequence depends entirely on the way you converted your DNA sequence into triplets before using the genetic code. For instance, using the second position as starting point leads to

```
1- ATGGAAGTATTTAAAGCGCCACCTATTGGGATATAAG

2- A TGG AAG TAT TTA AAG CGC CAC CTA TTG GGA TAT AAG

3-  W K Y L K R H L L G Y K
```

Beginning with the third position (GGA-AGT- . . .) again leads to an entirely different translation.

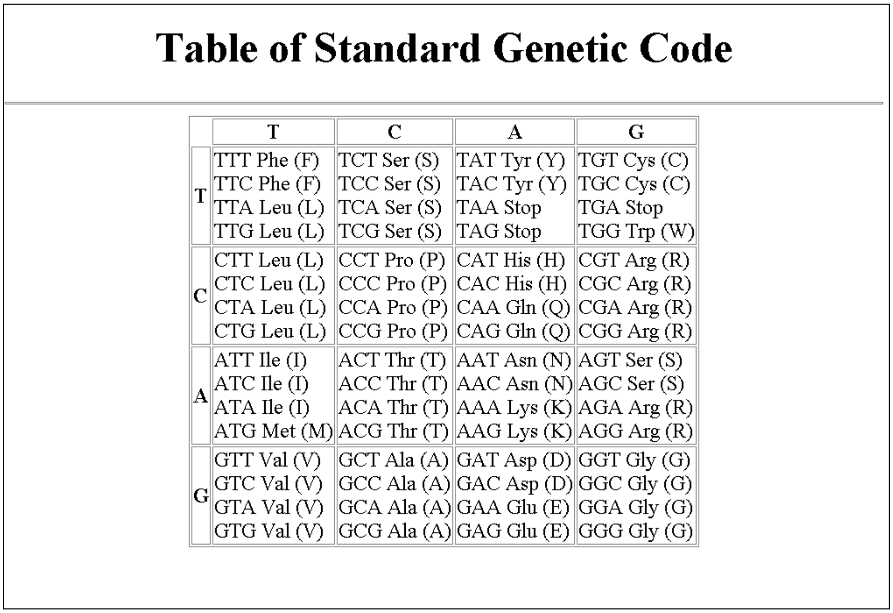


Table of Standard Genetic Code

	T	C	A	G
T	TTT Phe (F) TTC Phe (F) TTA Leu (L) TTG Leu (L)	TCT Ser (S) TCC Ser (S) TCA Ser (S) TCG Ser (S)	TAT Tyr (Y) TAC Tyr (Y) TAA Stop TAG Stop	TGT Cys (C) TGC Cys (C) TGA Stop TGG Trp (W)
C	CTT Leu (L) CTC Leu (L) CTA Leu (L) CTG Leu (L)	CCT Pro (P) CCC Pro (P) CCA Pro (P) CCG Pro (P)	CAT His (H) CAC His (H) CAA Gln (Q) CAG Gln (Q)	CGT Arg (R) CGC Arg (R) CGA Arg (R) CGG Arg (R)
A	ATT Ile (I) ATC Ile (I) ATA Ile (I) ATG Met (M)	ACT Thr (T) ACC Thr (T) ACA Thr (T) ACG Thr (T)	AAT Asn (N) AAC Asn (N) AAA Lys (K) AAG Lys (K)	AGT Ser (S) AGC Ser (S) AGA Arg (R) AGG Arg (R)
G	GTT Val (V) GTC Val (V) GTA Val (V) GTG Val (V)	GCT Ala (A) GCC Ala (A) GCA Ala (A) GCG Ala (A)	GAT Asp (D) GAC Asp (D) GAA Glu (E) GAG Glu (E)	GGT Gly (G) GGC Gly (G) GGA Gly (G) GGG Gly (G)

Figure 1-9: The universal genetic code.

Because of the triplet-based genetic code, a given DNA interval, *on a given strand,* can theoretically be translated in three different ways — basically three perspectives that are known in the field as *reading frames.* Because the DNA can be used from both strands, a total of six possible reading frames are possible for translating a DNA sequence into proteins. With very few exceptions (found in exotic viruses), only one of these six frames is used for any given DNA coding region. An interval of DNA sequence that remains free of STOP (the translation of TAA, TGA, or TAG) is called an *open reading frame* (ORF).

Additional complications arise from the fact that some DNA sequences are not encoding proteins at all — and that higher organisms have large pieces of noncoding DNA inserted within their genes. A large part of bioinformatics is devoted to the development of methods to locate protein-coding regions in DNA sequences, to delineate precisely where genes start and end, or where they are interrupted by the noncoding intervals (called *introns*).

DNA/RNA bioinformatics covered in this book

Need a road map to the bioinformatic analyses that are relevant to DNA/RNA sequences covered in this book? Here it is:

- ✓ Retrieving DNA sequences from databases (Chapters 2 and 3)
- ✓ Computing nucleotide compositions (Chapter 5)
- ✓ Identifying restriction sites (Chapter 5)
- ✓ Designing polymerase chain-reaction (PCR) primers (Chapter 5)
- ✓ Identifying open reading frames (ORFs) (Chapter 5)
- ✓ Predicting elements of DNA/RNA secondary structure (Chapter 12)
- ✓ Finding repeats (Chapter 5)
- ✓ Computing the optimal alignment between two or more DNA sequences (Chapters 7, 8, and 9)
- ✓ Finding polymorphic sites in genes (single nucleotide polymorphisms, SNPs) (Chapter 3)
- ✓ Assembling sequence fragments (Chapter 5)

Working with Entire Genomes

The first truly efficient technique to sequence DNA was discovered in 1977. In 1995, the first sequence of an entire genome (from the microbe *Hemophilus influenzae*) was determined. Between these two dates, DNA-sequencing technologies improved steadily, but such technologies still tended to concentrate on mining individual genes for information. During this period, biologists were mostly sequencing DNA fragments that were a few thousand nucleotides in length, simply because they were interested in specific genes that they had started working on years before. Most of the bioinformatics tools available today were created during that period. They include

- All basic sequence-alignment programs
- Phylogenetic and classification methods
- Various display tools adapted to relatively small-sequence objects (such as protein sequences no more than a few thousand characters long)

Genomics: Getting all the genes at once

The determination of the first complete genome sequence terminated the gene-by-gene routine and initiated the era of *genomics,* the genetic mapping, physical mapping, and sequencing of entire genomes. As a consequence, the DNA sequences we have to work with now are much longer — close to a million-bp in length for microbes and up to several billion-bp in length for animals and humans. This revolution called for the design of new bioinformatic tools and databases capable to store, query, analyze, and display these huge objects in a user-friendly manner. Chapters 3, 5, and 7 present some of the questions that biologists address at the genome scale, and show the relevant bioinformatic tools in action.

In contrast to the early days of the gene-by-gene approach, DNA sequences are now often obtained (along with the presumed protein sequences derived from those DNA sequences) without any prior knowledge of what is actually there. In essence, genes are both sequenced *and discovered* at the same time. This development prompted the emergence of an entirely new branch of bioinformatics devoted to the *parsing* of large DNA sequences into their components (genes, transcription units, protein-coding regions, regulatory elements, and so forth). This first pass is then followed by a longer phase of genome *annotation,* where the biological functions of these various elements are (more or less tentatively) predicted. Part IV of this book presents you with some of these most advanced techniques.

Figure 1-10, representing the whole genome of the bacterium *Rickettsia conorii,* illustrates this new level of complexity. This circular DNA molecule is 1.3 million bp long, on the small side for a bacterium. Each little rectangle in the two most external circles of features (one circle per strand) corresponds to a protein-coding gene in the circular genome. Each rectangle corresponds to approximately 1000 bp. Nobody knew which genes — or which proteins — were in that bacterium before the sequencing started. Almost everything we know now about this bacterium (and many others we can describe as fairly inaccessible, such as those thriving on the ocean floor near volcanic vents at 100°C) has been derived from bioinformatic analyses.

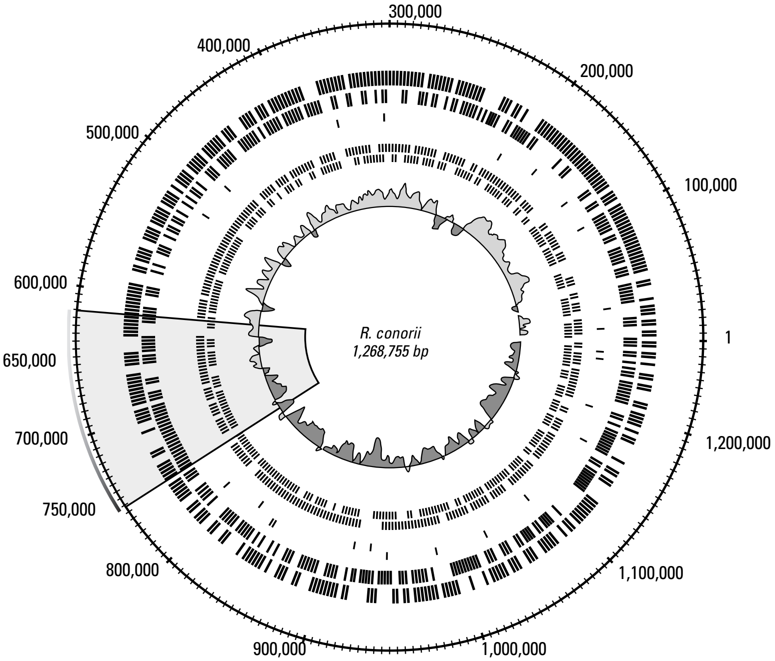


Figure 1-10: Representation of a bacterial genome.

Genome bioinformatics covered in this book

The following list lets you know where in this book you'll find more in-depth coverage of specific topics (some of them bristling with scary, mouth-filling terms) related to genome bioinformatics:

- Finding which genomes are available (Chapter 3)
- Analyzing sequences in relation to specific genomes (Chapters 3 and 7)
- Displaying genomes (Chapter 3)
- Parsing a microbial genome sequence: ORFing (Chapter 5)
- Parsing a eukaryotic genome sequence: GenScan (Chapter 5)
- Finding orthologous and paralogous genes (Chapter 3)
- Finding repeats (Chapter 5)

Chapter 2

How Most People Use Bioinformatics

In This Chapter

- ▶ Using PubMed to become quickly knowledgeable on any biological subject
- ▶ Retrieving protein and DNA sequences relevant to your work
- ▶ Running your first sequence/database comparison with BLAST
- ▶ Performing your first multiple sequence alignment with ClustalW

> *Personally, I'm always ready to learn, although I do not always like being taught.*
>
> — Sir Winston Churchill

In this chapter, we show you enough bioinformatics to perform the routine — but nonetheless incredibly useful — tasks that experienced molecular biologists once spent months mastering by wandering helplessly on the Net. Chances are that you can fulfill 90 percent of your professional needs without knowing more than the content of this chapter. If you've already got a pretty good chunk of bioinformatics under your belt — enough to qualify as a power user — you may still find useful reminders and tricks here. If your background is in computer science, we show you the kinds of tools that biologists demand. So get on your favorite PC, follow the guide, and enjoy the ride!

Becoming an Instant Expert with PubMed/Medline

Think about the following situation: You just got the sequence of the gene you've been working on for years, and you gave it to the grumpy local specialist in bioinformatics to analyze it for you. (This is, of course, *before* you had an opportunity to read this book!) The specialist's diagnosis comes back, and here it is: "Jim, your sequence looks pretty much like a dUTPase."

Finding out about a protein by its name

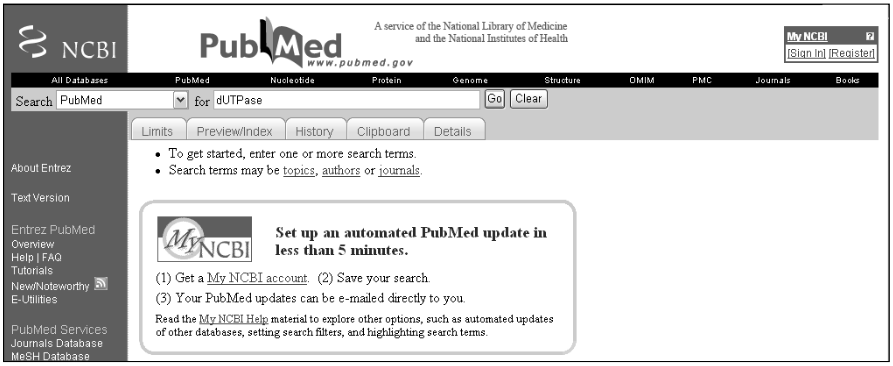
The good news is, your gene actually resembles *something,* and you may have the beginning of a *story.* The bad news is that you've never heard the term "dUTPase" in your entire life. Where do you go from here? How do you find out more about this seemingly obscure term with the strange spelling? How do you quickly decide whether this story makes any sense in the context of what you already know about your gene?

Searching PubMed

The answer is to stay in your chair and go to PubMed! To do so, follow these steps:

1. **Using your favorite Internet browser, navigate to** `www.ncbi.nlm.nih.gov/entrez/` **on the World Wide Web.**

   Your screen now looks like Figure 2-1, with a Search window already preset to PubMed and a For text box next to it, ready for you to type in a few words relevant to your question.



Search PubMed for dUTPase

- To get started, enter one or more search terms.
- Search terms may be topics, authors or journals.

Set up an automated PubMed update in less than 5 minutes.

(1) Get a My NCBI account. (2) Save your search.

(3) Your PubMed updates can be e-mailed directly to you.

Read the My NCBI Help material to explore other options, such as automated updates of other databases, setting search filters, and highlighting search terms.

Figure 2-1: The initial PubMed search window.

2. **Type in** dUTPase **in the For window, and click the Go button.**

   Soon after, a Results list like the one in Figure 2-2 appears on-screen.

   For our dUTPase example, we now have more than 200 references at our fingertips, more than enough to start unraveling the mysteries of dUTPase — including its relevance to your gene. But don't feel you have to rush to the library with your printed list of references before it closes: You may not even have to go.

3. **For any entry in the Results list, click the associated author names.**

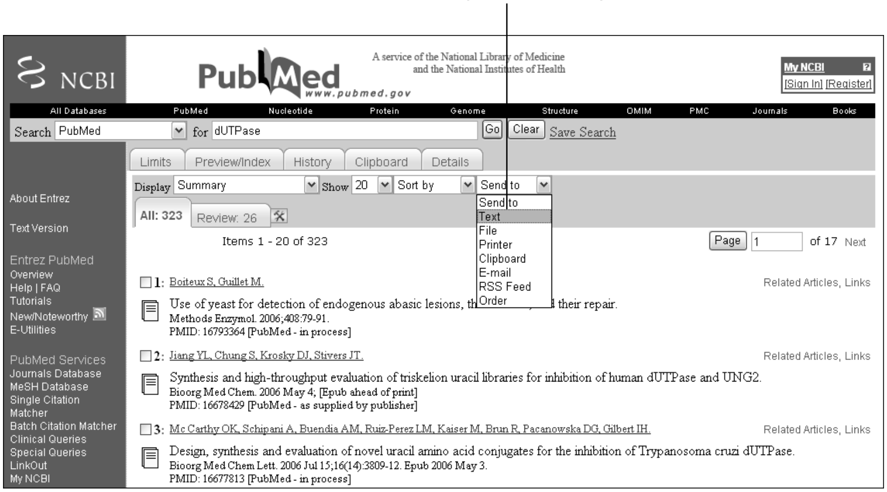
   Author names stand out because they appear in a blue font, underlined — two signs of a clickable hyperlink.

Clicking the link calls up a page containing a rather detailed summary of the paper you chose. If you have a reasonable background in biology and a little academic flair, reading a few of these abstracts should be enough for you to get an idea of what dUTPase is all about.

4. **Save what you like to your hard drive by choosing your browser's File➪Save As option.**

   Alternatively, you can transform the display into a simpler (more printer-friendly) format by selecting Text from the Send To drop-down menu and then choosing File➪Save As.

Click here for printer-friendly format.



Search PubMed for dUTPase

Display Summary Show 20 Sort by Send to

All: 323 Review: 26

Items 1 - 20 of 323

Page 1 of 17

□ 1: Boiteux S, Guillet M.

Use of yeast for detection of endogenous abasic lesions, th their repair.
Methods Enzymol. 2006;408:79-91.
PMID: 16793364 [PubMed - in process]

□ 2: Jiang YL, Chung S, Krosky DJ, Stivers JT.

Synthesis and high-throughput evaluation of triskelion uracil libraries for inhibition of human dUTPase and UNG2.
Bioorg Med Chem. 2006 May 4; [Epub ahead of print]
PMID: 16678429 [PubMed - as supplied by publisher]

□ 3: Mc Carthy OK, Schipani A, Buendia AM, Ruiz-Perez LM, Kaiser M, Brun R, Pacanowska DG, Gilbert IH.

Design, synthesis and evaluation of novel uracil amino acid conjugates for the inhibition of Trypanosoma cruzi dUTPase.
Bioorg Med Chem Lett. 2006 Jul 15;16(14):3809-12. Epub 2006 May 3.
PMID: 16677813 [PubMed - in process]

Figure 2-2: The initial result of a standard PubMed search for dUTPase.

Saving multiple summaries

When your search yields many references, the best move is to start scanning a few pages — and select the most promising papers for future use by checking the corresponding boxes. To move through the Results pages, just use the Next link (top right).

When you think you have selected enough references, just do the following:

1. **Choose Abstract from the Display drop-down menu.**

   The abstracts of all the papers you selected in the various pages appear for you to read.

2. **If you want to print this display as is, choose File➪Print from your browser menu.**

3. **If you want to print this display in a more printer-friendly format, first select Text from the scroll-down menu on the far right of the menu bar (refer to Figure 2-2) to display this information in a non-HTML format, and then choose File➪Print from your browser menu.**
4. **If you'd like to save the file in the format of your choice, choose File➪Save As from your browser menu, enter a new filename for the file, and then choose the file format you want to save to.**

TIP

Although this particular drop-down menu — the one you see fully extended back in Figure 2-2 — provides a File option, it sometimes conflicts with security settings. We advise you to stay away from it and use the standard File➪Save As command on your browser menu.

Searching PubMed using author's names

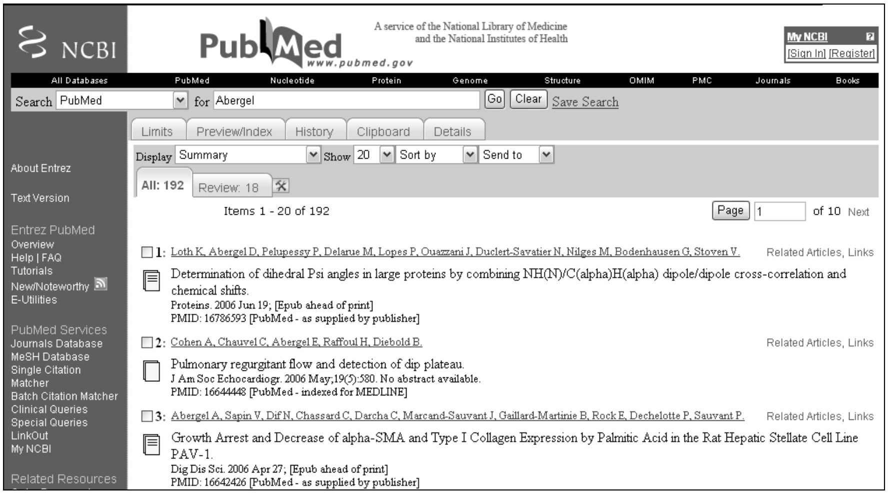
Although you can certainly search PubMed by using topic keywords (such as protein names), you do have other search options available to you. For example, we're sure you've found yourself in the situation where one of your colleagues (or perhaps your advisor) has told you something like the following: "I read a paper by so and so, published not too long ago, and I think it was on dUTPase; you should check it out." In the old-style academic world of card catalogs and stacks of periodicals, this isn't much of a lead. But on the World Wide Web, with PubMed at your side, you're almost home free. Suppose you only remember one of these author's names, such as Abergel. Using PubMed, take the following steps to track down the elusive reference:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`.
2. **Type** Abergel **— the name of your prospective author — in the For window, and then click the Go button.**

   The Results list appears, as shown in Figure 2-3.

   We now have a list of many papers, all of them with "Abergel" as an author. Again, you can browse through this list, select some of them, and bring out their abstracts. However, the bad news is that none of them appear to have anything to do with dUTPase. To solve this problem, you simply combine the protein name and the author information in your query — and you don't even need to memorize a complicated syntax or Boolean symbols.
3. **Type** dUTPase **next to** Abergel **in the For window — don't forget to put a single space between the search terms — and click Go.**

   By default, PubMed assumes that you want to use both search terms — `Abergel` AND `dUTPase` — when looking for the appropriate research papers. By default, PubMed also assumes that you want to look for these words in the titles OR in the abstracts of the papers. Figure 2-4 shows the results of your first combined search.



Search PubMed for Abergel

Display Summary Show 20 Sort by Send to

All: 192 Review: 18

Items 1 - 20 of 192 Page 1 of 10 Next

□ 1: Loth K, Abergel D, Pelupessy P, Delarue M, Lopes P, Ouazzani J, Duclert-Savatier N, Nilges M, Bodenhausen G, Stoven V. Related Articles, Links

Determination of dihedral Psi angles in large proteins by combining NH(N)/C(alpha)H(alpha) dipole/dipole cross-correlation and chemical shifts.
Proteins. 2006 Jun 19; [Epub ahead of print]
PMID: 16786593 [PubMed - as supplied by publisher]

□ 2: Cohen A, Chauvel C, Abergel E, Raffoul H, Diebold B. Related Articles, Links

Pulmonary regurgitant flow and detection of dip plateau.
J Am Soc Echocardiogr. 2006 May;19(5):580. No abstract available.
PMID: 16644448 [PubMed - indexed for MEDLINE]

□ 3: Abergel A, Sapin V, Dif N, Chassard C, Darcha C, Marcand-Sauvant J, Gaillard-Martinie B, Rock E, Dechelotte P, Sauvant P. Related Articles, Links

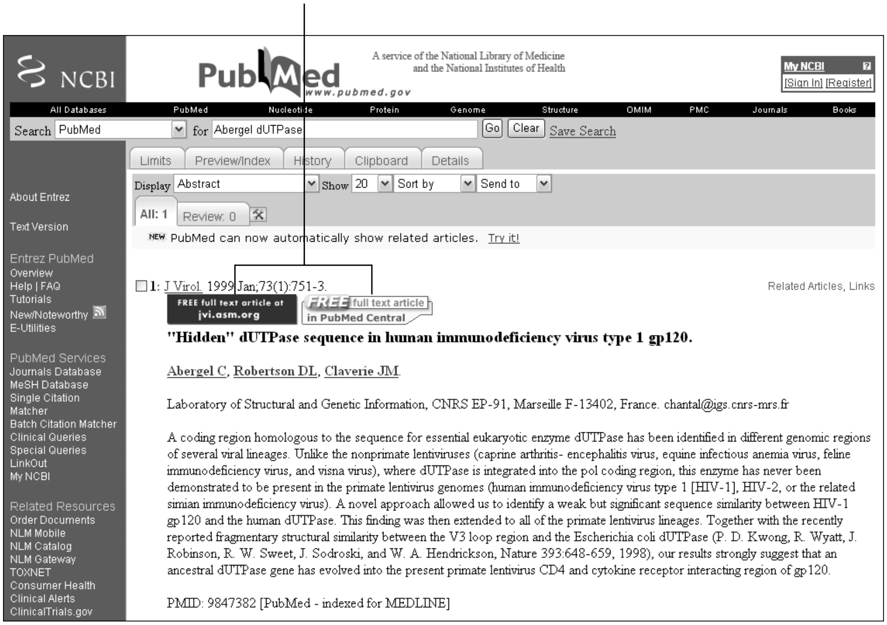
Growth Arrest and Decrease of alpha-SMA and Type I Collagen Expression by Palmitic Acid in the Rat Hepatic Stellate Cell Line PAV-1.
Dig Dis Sci. 2006 Apr 27; [Epub ahead of print]
PMID: 16642426 [PubMed - as supplied by publisher]

Figure 2-3: The results of a standard PubMed search for Abergel.

Because only one paper corresponds to your query, PubMed automatically switches from the "summary" to the "abstract" mode, and you can now read what this unique paper is all about.

This must be the paper your boss was talking about. Congratulations! You've just accomplished your first bibliographic search with PubMed!

Click here for full text.



Search PubMed for Abergel dUTPase

Display Abstract Show 20 Sort by Send to

All: 1 Review: 0

NEW PubMed can now automatically show related articles. Try it!

□ 1: J Virol. 1999 Jan;73(1):751-3. Related Articles, Links

"Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.

Abergel C, Robertson DL, Claverie JM.

Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-13402, France. chantal@igs.cnrs-mrs.fr

A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. Unlike the nonprimate lentiviruses (caprine arthritis- encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus), where dUTPase is integrated into the pol coding region, this enzyme has never been demonstrated to be present in the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related simian immunodeficiency virus). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1 gp120 and the human dUTPase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently reported fragmentary structural similarity between the V3 loop region and the Escherichia coli dUTPase (P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

PMID: 9847382 [PubMed - indexed for MEDLINE]

Figure 2-4: The result of a standard combined search for Abergel and dUTPase.

When a search returns more than a single reference, simply choose the Abstract option manually from the Display drop-down menu.

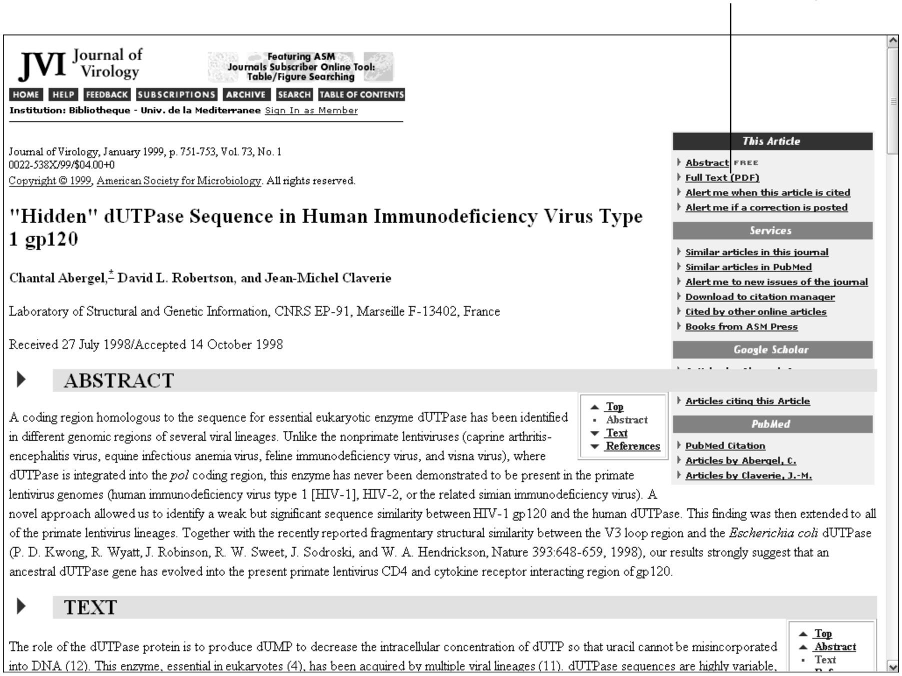
But wait! There's more! In fact, this is your lucky day: You've hit upon a paper from a journal that offers free access to the full text (as indicated by the large colored rectangles that appear above the title).

4. **Click the blue rectangle on the left.**

   This brings you directly to the publisher's site, where you read an interactive online copy (HTML) or download a reprint in .PDF format, among other choices (Figure 2-5). An increasing number of scientific journals — including some leading journals in their respective fields — are offering full-text access to articles. However, most of them will still ask you to pay a fee, to be a registered client (for instance, through your university), or to hold a personal subscription. For some journals (such as the one in Figure 2-5), full-text access becomes free six months after print publication.

   By the way, clicking on the rectangle next to the blue one gives you access to the same article, this time extracted from the public, open-access bibliographic repository PubMed Central.

   Thus, if you're lucky, you can go through the entire bibliographic process in a flash — and find out all you need about some mysterious biological subject — without having to leave the comfort of your chair, thanks to PubMed and the cooperating publishers.


Click here for a PDF reprint.

JVI Journal of Virology
Featuring ASM Journals Subscriber Online Tool: Table/Figure Searching
HOME
HELP
FEEDBACK
SUBSCRIPTIONS
ARCHIVE
SEARCH
TABLE OF CONTENTS
This Article
Abstract FREE
Full Text (PDF)
Alert me when this article is cited
Alert me if a correction is posted
Services
Similar articles in this journal
Similar articles in PubMed
Alert me to new issues of the journal
Download to citation manager
Cited by other online articles
Books from ASM Press
Google Scholar
Articles citing this Article
PubMed
PubMed Citation
Articles by Abergel, C.
Articles by Claverie, J.-M.
Top
Abstract
Text
References

Institution: Bibliotheque - Univ. de la Mediterranee Sign In as Member

Journal of Virology, January 1999, p. 751-753, Vol. 73, No. 1
0022-538X/99/$04.00+0
Copyright © 1999, American Society for Microbiology. All rights reserved.

"Hidden" dUTPase Sequence in Human Immunodeficiency Virus Type 1 gp120

Chantal Abergel,[*] David L. Robertson, and Jean-Michel Claverie

Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille F-13402, France

Received 27 July 1998/Accepted 14 October 1998

ABSTRACT

A coding region homologous to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions of several viral lineages. Unlike the nonprimate lentiviruses (caprine arthritis-encephalitis virus, equine infectious anemia virus, feline immunodeficiency virus, and visna virus), where dUTPase is integrated into the *pol* coding region, this enzyme has never been demonstrated to be present in the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related simian immunodeficiency virus). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1 gp120 and the human dUTPase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently reported fragmentary structural similarity between the V3 loop region and the *Escherichia coli* dUTPase (P. D. Kwong, R. Wyatt, J. Robinson, R. W. Sweet, J. Sodroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an ancestral dUTPase gene has evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

TEXT

The role of the dUTPase protein is to produce dUMP to decrease the intracellular concentration of dUTP so that uracil cannot be misincorporated into DNA (12). This enzyme, essential in eukaryotes (4), has been acquired by multiple viral lineages (11). dUTPase sequences are highly variable,

Figure 2-5: The Journal of Virology grants free full-text access.

PubMed searches are *case-insensitive;* the results of the searches that appear in Figures 2-1 through 2-4 do not depend on the capitalization of query terms when you type them. Thus typing *DUTPASE, dUTPase,* or *dutpase* would produce the same output (as would typing *Abergel* or *aBERGEL*). Be aware, however, that this is NOT the case with all databases and search programs used in bioinformatics. Watching for case sensitivity is always a good idea.

Searching PubMed using fields

In addition to searching by author and topic, PubMed offers you many more ways to narrow down your bibliographic searches to more specific topics. To get a handle on these different ways, however, we need to know more about the internal structure of Medline entries — the pool of entries from which PubMed gathers its search results.

To gain some insight into this internal structure, refer to Figure 2-4. We had just used the PubMed site (`www.ncbi.nlm.nih.gov/entrez/`) to track down a single article after inputting *Abergel* and *dUTPase* as a query. With that Results list on your computer screen, you can now do the following:

1. **Click the small arrow to the right of the Display drop-down menu.**

   The contents of the drop-down menu appear: These options are different ways to display information related to the current article, or are links to related information.

2. **Select the MEDLINE option (Figure 2-6).**

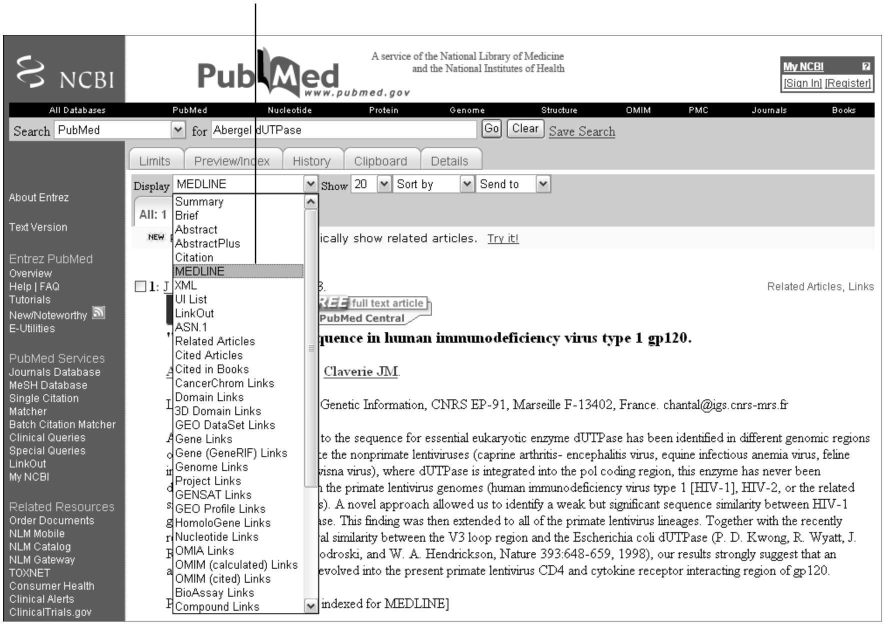
   The Medline page appears, as shown in Figure 2-7.

In Figure 2-7 you can see the internal structure of a Medline database record. The information is spread out over separate sections, called *fields,* each one preceded by a specific abbreviation — `TI` for title fields, `AB` for abstracts, `AD` for the laboratory address, `AU` for the authors, `SO` for the journal abbreviation, and so forth. This structure applies to all Medline records.

Putting fields to good use

Unfortunately, not all PubMed searches are as easy and productive as the one we used in the dUTPase example. You can get flooded with an overwhelming number of *hits* (the standard term for search results) if you formulate queries that contain common names (such as *Smith* or *Cohen*) or use search terms (such as *Down*) that can occur in different contexts — such as titles (for example, *Down Syndrome*), abstracts, author names, or even addresses (*955 Down Street*).

Select MEDLINE format here.



Search PubMed for Abergel dUTPase

Display MEDLINE Show 20 Sort by Send to

ically show related articles. Try it!

1: J 3.

quence in human immunodeficiency virus type 1 gp120.

Claverie JM.

Genetic Information, CNRS EP-91, Marseille F-13402, France. chantal@igs.cnrs-mrs.fr

to the sequence for essential eukaryotic enzyme dUTPase has been identified in different genomic regions
e the nonprimate lentiviruses (caprine arthritis- encephalitis virus, equine infectious anemia virus, feline
visna virus), where dUTPase is integrated into the pol coding region, this enzyme has never been
 the primate lentivirus genomes (human immunodeficiency virus type 1 [HIV-1], HIV-2, or the related
s). A novel approach allowed us to identify a weak but significant sequence similarity between HIV-1
ase. This finding was then extended to all of the primate lentivirus lineages. Together with the recently
al similarity between the V3 loop region and the Escherichia coli dUTPase (P. D. Kwong, R. Wyatt, J.
odroski, and W. A. Hendrickson, Nature 393:648-659, 1998), our results strongly suggest that an
evolved into the present primate lentivirus CD4 and cytokine receptor interacting region of gp120.

indexed for MEDLINE]

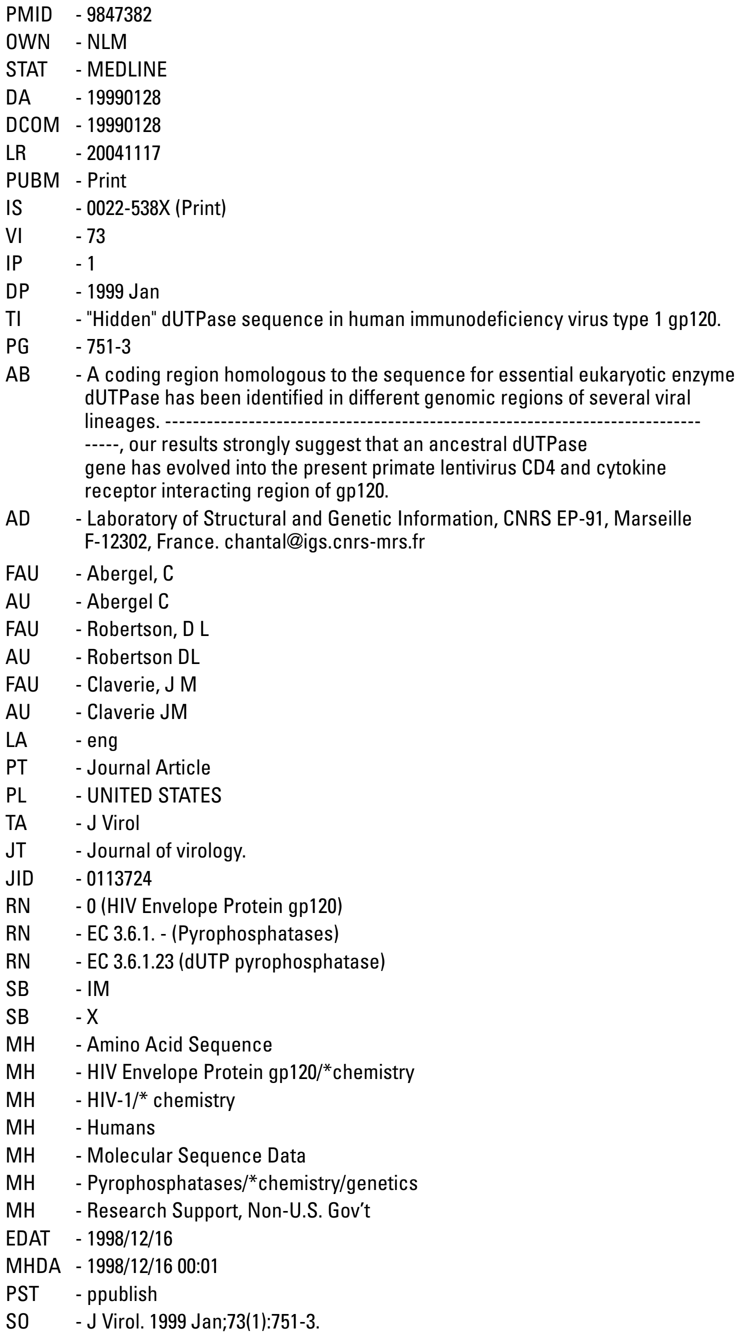
Figure 2-6: Changing the default abstract format to MEDLINE.

To alleviate this problem, you can query PubMed by restricting the search for each word of your query to a given field. Thus papers containing the word at the wrong place are not selected. To do so, you simply have to follow each term with the code (in brackets) that identifies the field in which you want to find the search term. Changing the field can totally modify the result of the searches. For instance, try entering three different queries — `Down [AU]`, `Down [TI]`, and `Down [AD]` into the For text box at the PubMed site.

When we ran those queries, we got (respectively) 275, 16,318, and 1,213 totally unrelated references. At least that narrowed the search a bit.

Using fields to find experts near you

Field-restricted queries are great for dealing with author's names or subject terms that can be found in different fields (which would otherwise inflate and confuse the search results). But field-restricted queries have many other applications, as well — they're especially useful for locating experts in particular research fields in given locations.



```
PMID    - 9847382
OWN     - NLM
STAT    - MEDLINE
DA      - 19990128
DCOM    - 19990128
LR      - 20041117
PUBM    - Print
IS      - 0022-538X (Print)
VI      - 73
IP      - 1
DP      - 1999 Jan
TI      - "Hidden" dUTPase sequence in human immunodeficiency virus type 1 gp120.
PG      - 751-3
AB      - A coding region homologous to the sequence for essential eukaryotic enzyme
          dUTPase has been identified in different genomic regions of several viral
          lineages. --------------------------------------------------------------------------
          -----, our results strongly suggest that an ancestral dUTPase
          gene has evolved into the present primate lentivirus CD4 and cytokine
          receptor interacting region of gp120.
AD      - Laboratory of Structural and Genetic Information, CNRS EP-91, Marseille
          F-12302, France. chantal@igs.cnrs-mrs.fr
FAU     - Abergel, C
AU      - Abergel C
FAU     - Robertson, D L
AU      - Robertson DL
FAU     - Claverie, J M
AU      - Claverie JM
LA      - eng
PT      - Journal Article
PL      - UNITED STATES
TA      - J Virol
JT      - Journal of virology.
JID     - 0113724
RN      - 0 (HIV Envelope Protein gp120)
RN      - EC 3.6.1. - (Pyrophosphatases)
RN      - EC 3.6.1.23 (dUTP pyrophosphatase)
SB      - IM
SB      - X
MH      - Amino Acid Sequence
MH      - HIV Envelope Protein gp120/*chemistry
MH      - HIV-1/* chemistry
MH      - Humans
MH      - Molecular Sequence Data
MH      - Pyrophosphatases/*chemistry/genetics
MH      - Research Support, Non-U.S. Gov't
EDAT    - 1998/12/16
MHDA    - 1998/12/16 00:01
PST     - ppublish
SO      - J Virol. 1999 Jan;73(1):751-3.
```

Figure 2-7: The internal structure of a Medline record.

Suppose (for example) you just landed in Chicago and you'd like to know if there's anybody around who can give you some insight into the mysteries of dUTPase. To get your personalized list of resident experts, just do the following:

1. **Point your browser to the PubMed site** (`www.ncbi.nlm.nih.gov/entrez/`).
2. **In the For window, enter** dUTPase [TIAB] Chicago [AD]**, and click Go.**

   By specifying the `[TIAB]` field, you'll be scanning the titles and abstracts of potential articles; the `[AD]` field specifies the address of the main laboratory associated with these articles.

   A couple of papers show up (Figure 2-8).
3. **Click the list of authors (in blue, underlined).**

   You get the abstract of the corresponding article, together with the main laboratory address. Alternatively, to get all the abstracts at once, follow Steps 4 and 5.
4. **Go back to the previous URL (after Step 2) using the Go Back button of your browser.**
5. **In the Display drop-down menu, change the display option from *Summary* to *Abstract.***

   A list of abstracts appears. At the top of each abstract, you get the name of the relevant laboratory in Chicago.

   With the information contained in the Medline records, you have names, street addresses, and sometimes e-mail addresses. If necessary, you can use a telephone book (or a Web search engine) to supplement this information and find out how to contact these experts — the same day. (But please don't call these people on our behalf!)

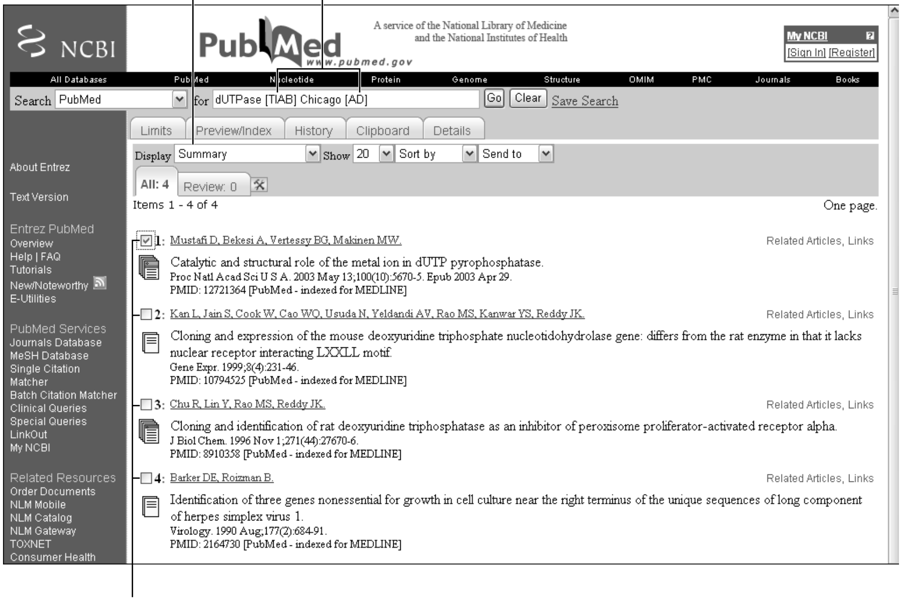
Searching PubMed using limits

PubMed offers yet another way of fine-tuning your queries: You can pre-define ranges for different attributes in different fields before you run your search. Said like this, the process sounds awfully technical and complex — but here's where we show you how to do it (again using our dUTPase example).

When you want to quickly find out the basics about a subject you know nothing about, it's best not to read any single highly specialized article. A better bet is to stick with review articles, where one expert summarizes the state of the art for you. Unless you're an historian, you won't have much need to know about the state of the art as it was 20 years ago (which seems more like three centuries ago in the field of molecular biology). Thus, we want to limit our PubMed search to *recent* review articles about dUTPase. Here's how it's done:

Change from Summary to abstract here.

Restricted fields



Search PubMed for dUTPase [TIAB] Chicago [AD]

Display Summary Show 20

Items 1 - 4 of 4 One page.

☑ 1: Mustafi D, Bekesi A, Vertessy BG, Makinen MW. Related Articles, Links
Catalytic and structural role of the metal ion in dUTP pyrophosphatase.
Proc Natl Acad Sci U S A. 2003 May 13;100(10):5670-5. Epub 2003 Apr 29.
PMID: 12721364 [PubMed - indexed for MEDLINE]

□ 2: Kan L, Jain S, Cook W, Cao WQ, Usuda N, Yeldandi AV, Rao MS, Kanwar YS, Reddy JK. Related Articles, Links
Cloning and expression of the mouse deoxyuridine triphosphate nucleotidohydrolase gene: differs from the rat enzyme in that it lacks nuclear receptor interacting LXXLL motif.
Gene Expr. 1999;8(4):231-46.
PMID: 10794525 [PubMed - indexed for MEDLINE]

□ 3: Chu R, Lin Y, Rao MS, Reddy JK. Related Articles, Links
Cloning and identification of rat deoxyuridine triphosphatase as an inhibitor of peroxisome proliferator-activated receptor alpha.
J Biol Chem. 1996 Nov 1;271(44):27670-6.
PMID: 8910358 [PubMed - indexed for MEDLINE]

□ 4: Barker DE, Roizman B. Related Articles, Links
Identification of three genes nonessential for growth in cell culture near the right terminus of the unique sequences of long component of herpes simplex virus 1.
Virology. 1990 Aug;177(2):684-91.
PMID: 2164730 [PubMed - indexed for MEDLINE]

Check these boxes to select papers.

Figure 2-8: Searching for experts on dUTPase in Chicago.

1. **Point your browser to the PubMed site at**

   `www.ncbi.nlm.nih.gov/entrez/`

2. **Type** dUTPase **in the For text box.**

   If you clicked the Go button at this point, you'd get a list of more than 320 articles — a bit too much reading for a nonspecialist! Restricting this list to the most recent review articles written in English would be practical.

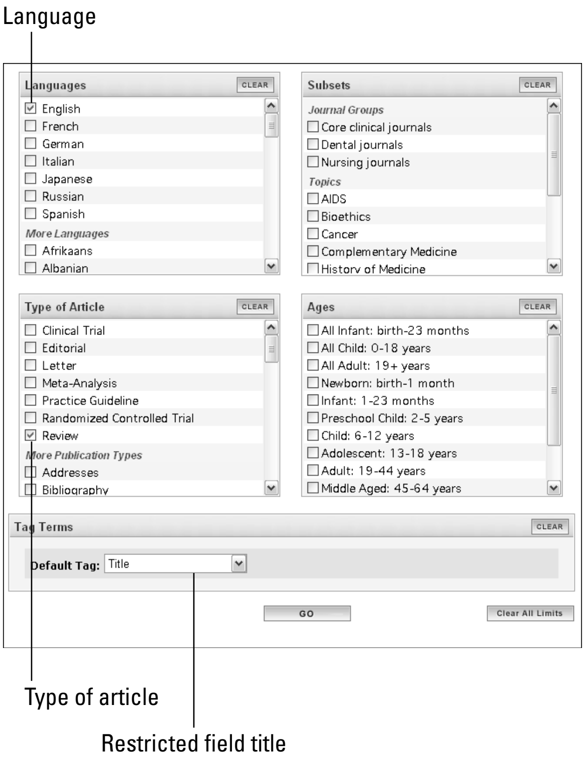
3. **Click the Limits tab, located just beneath the little arrow for the pull-down menu of the Search window.**

   The Limits screen appears, as shown in Figure 2-9. Here, below the Limited To line, you now have plenty of fields and attributes you can use for setting limits. You can go back later to this page and explore these various options.

4. **Check the English box in the Language section.**

   Unless you're fluent in French, *bien sûr.*

5. **Check the Review box in the Type of Article section.**



Languages CLEAR
☑ English
☐ French
☐ German
☐ Italian
☐ Japanese
☐ Russian
☐ Spanish
More Languages
☐ Afrikaans
☐ Albanian

Subsets CLEAR
Journal Groups
☐ Core clinical journals
☐ Dental journals
☐ Nursing journals
Topics
☐ AIDS
☐ Bioethics
☐ Cancer
☐ Complementary Medicine
☐ History of Medicine

Type of Article CLEAR
☐ Clinical Trial
☐ Editorial
☐ Letter
☐ Meta-Analysis
☐ Practice Guideline
☐ Randomized Controlled Trial
☑ Review
More Publication Types
☐ Addresses
☐ Bibliography

Ages CLEAR
☐ All Infant: birth-23 months
☐ All Child: 0-18 years
☐ All Adult: 19+ years
☐ Newborn: birth-1 month
☐ Infant: 1-23 months
☐ Preschool Child: 2-5 years
☐ Child: 6-12 years
☐ Adolescent: 13-18 years
☐ Adult: 19-44 years
☐ Middle Aged: 45-64 years

Tag Terms CLEAR
Default Tag: Title

GO Clear All Limits

Figure 2-9: Limiting a search for dUTPase to the titles of review articles in English.

6. **Choose Title from the Default Tag drop-down menu.**

   Choosing Title restricts the search to articles for which the topic of dUTPase is central.

7. **Finally, click the Go button.**

   Your (new and more concise) Results page appears.

In our hands, searching for dUTPase using these limits resulted in five articles — much better than the 200 hits we would have gotten without the limits. More important, we've limited the search to *review* articles, which hopefully contain — in concise form — all we'd ever want to know about this subject.

TIP

A common use of the Limit menu is also to restrict the search for papers published within a given date range (such as the most recent).

A few more tips about PubMed

Finding out how to use PubMed to its full extent is certainly worth your time because PubMed can be useful in many different ways. Remember that you cannot break anything, and you risk nothing by exploring what all these buttons — especially the ones we haven't talked about — might do. Here are a few last-minute tips:

- **How to get the most out of your query:**
  - Quoted queries (for example, `"down syndrome"`) behave as a single word, and are a great way to improve the relevance of your search.
  - Impress your colleagues by starting using logical connectors (`AND`, `OR`, `NOT`) in your queries, as in

    dUTPase[TI] OR pyrophosphatase[TI] NOT Smith[AU]
  - Adding initials to proper names (for example, `"Abergel C"`) can greatly reduce the number of hits.
  - Write down the PubMed Identifier (the number in the PMID field) of that interesting paper you just found. It can be very useful in any subsequent searches for related items, such as associated gene and protein sequences.
  - Don't forget to deselect the Limit box when starting a new search.
  - Don't put too much initial faith in a search that produces no results. Spelling mistakes, wrong field restrictions, or improper limits settings can all throw off an effective search.
  - As a beginner researching a new subject, read through a couple of abstracts to enlarge your initial "jargon" vocabulary — and look for synonyms. For example, if you don't know that some papers on dUTPase might use the term "dUTP pyrophosphatase" instead, you may miss out on some interesting papers.
  - Try the Related Articles link — the one to the extreme right of the PubMed output — to enlarge a search that isn't giving you enough references.
- **Things you (unfortunately) *won't* find in PubMed:**
  - **Names ranking beyond the 10th place in the author's list for older papers (before 1995).** This is a significant problem for many pioneering articles in genomics.
  - **Papers recorded before 1965.** PubMed doesn't have any. Don't rely on PubMed as a primary source if you're writing an historical article in your field.
  - **Abstracts for most references recorded before 1976.** Don't expect great results from PubMed searches involving these.

Retrieving Protein Sequences

If putting PubMed to use for bibliographic searches is the most common bioinformatics thing biologists want to do, then the *next* most popular thing is to retrieve relevant protein sequences from the Web to find out more about the subject at the molecular level. Although PubMed provides you with a direct link to the protein world, we find that it's sometimes confusing to use. So here we introduce you to another site, where finding protein sequences is really easy.

ExPASy: A prime Internet site for protein information

Created and managed by one of the pioneers of protein bioinformatics, Prof. Amos Bairoch, the ExPASy server is a world-leading resource for protein information. In addition to the Swiss-Prot protein sequence database, the ExPASy site provides numerous analysis tools that we use throughout this book. It also provides a wealth of outside links for more specialized analyses on other servers.

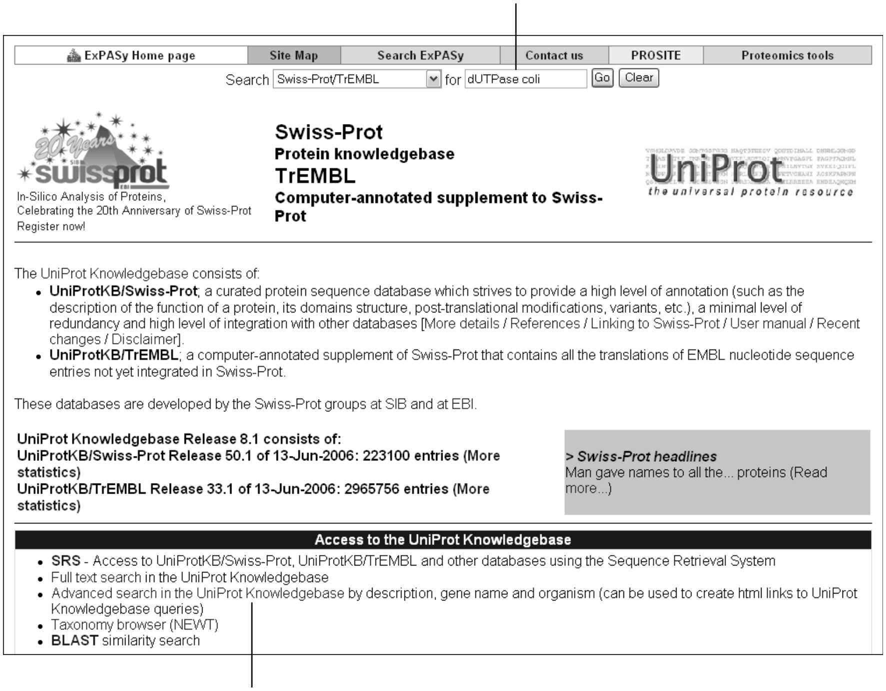
After using PubMed to acquire some preliminary information about a particular function that you're interested in — dUTPase, for example — go ahead and find out more about it by retrieving a few examples of protein sequences that perform this function. (Don't worry — we show you how.) Because *Escherichia coli* (abbreviated *E. coli*) is one of the most studied organisms, we use it in our example to show you how to retrieve the sequence of the protein performing the dUTPase function in this bacterium.

1. **Point your browser to** `www.expasy.org/sprot/`**, the Swiss-Prot database home page (Figure 2-10).**
2. **Type dUTPase coli in the Search window, and then click the Search button.**

   A list of three relevant protein sequences (as shown in Figure 2-11) appears.

   Now, when you click the last DUT_ECOLI (P06968) link, a full page of information about this dUTPase protein *of E. coli* appears on-screen, as shown in Figures 2-12 and 2-13. (No single browser window can hold such a wealth of information, so we've broken up the Results page into two figures.)

Search window



In-Silico Analysis of Proteins,
Celebrating the 20th Anniversary of Swiss-Prot
Register now!

Swiss-Prot
Protein knowledgebase
TrEMBL
Computer-annotated supplement to Swiss-Prot

The UniProt Knowledgebase consists of:

- **UniProtKB/Swiss-Prot**; a curated protein sequence database which strives to provide a high level of annotation (such as the description of the function of a protein, its domains structure, post-translational modifications, variants, etc.), a minimal level of redundancy and high level of integration with other databases [More details / References / Linking to Swiss-Prot / User manual / Recent changes / Disclaimer].
- **UniProtKB/TrEMBL**; a computer-annotated supplement of Swiss-Prot that contains all the translations of EMBL nucleotide sequence entries not yet integrated in Swiss-Prot.

These databases are developed by the Swiss-Prot groups at SIB and at EBI.

UniProt Knowledgebase Release 8.1 consists of:
UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006: 223100 entries (More statistics)
UniProtKB/TrEMBL Release 33.1 of 13-Jun-2006: 2965756 entries (More statistics)

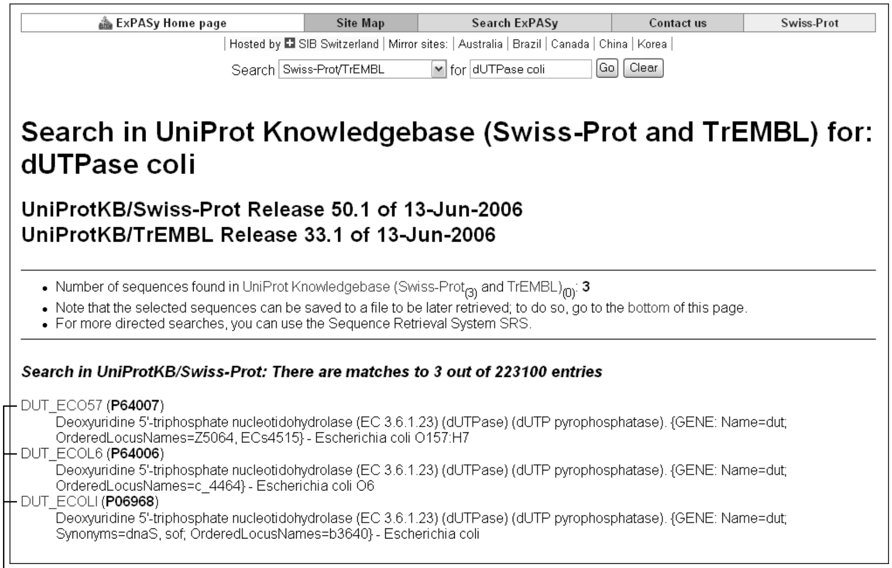
> Swiss-Prot headlines
Man gave names to all the... proteins (Read more...)

Access to the UniProt Knowledgebase

- **SRS** - Access to UniProtKB/Swiss-Prot, UniProtKB/TrEMBL and other databases using the Sequence Retrieval System
- Full text search in the UniProt Knowledgebase
- Advanced search in the UniProt Knowledgebase by description, gene name and organism (can be used to create html links to UniProt Knowledgebase queries)
- Taxonomy browser (NEWT)
- **BLAST** similarity search

Click here for Advanced Search.

Figure 2-10: The Search window at the top of the Swiss-Prot home page.



Search in UniProt Knowledgebase (Swiss-Prot and TrEMBL) for: dUTPase coli

UniProtKB/Swiss-Prot Release 50.1 of 13-Jun-2006
UniProtKB/TrEMBL Release 33.1 of 13-Jun-2006

- Number of sequences found in UniProt Knowledgebase (Swiss-Prot$_{(3)}$ and TrEMBL)$_{(0)}$: **3**
- Note that the selected sequences can be saved to a file to be later retrieved; to do so, go to the bottom of this page.
- For more directed searches, you can use the Sequence Retrieval System SRS.

Search in UniProtKB/Swiss-Prot: There are matches to 3 out of 223100 entries

DUT_ECO57 (**P64007**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; OrderedLocusNames=Z5064, ECs4515} - Escherichia coli O157:H7

DUT_ECOL6 (**P64006**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; OrderedLocusNames=c_4464} - Escherichia coli O6

DUT_ECOLI (**P06968**)
Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). {GENE: Name=dut; Synonyms=dnaS, sof; OrderedLocusNames=b3640} - Escherichia coli

3 relevant dUTPase entries found.

Figure 2-11: Three dUTPase sequences from the Escherichia coli bacterium.

TIP

A typical Swiss-Prot entry is made up of four parts. Here's how that looks for our example of dUTPase:

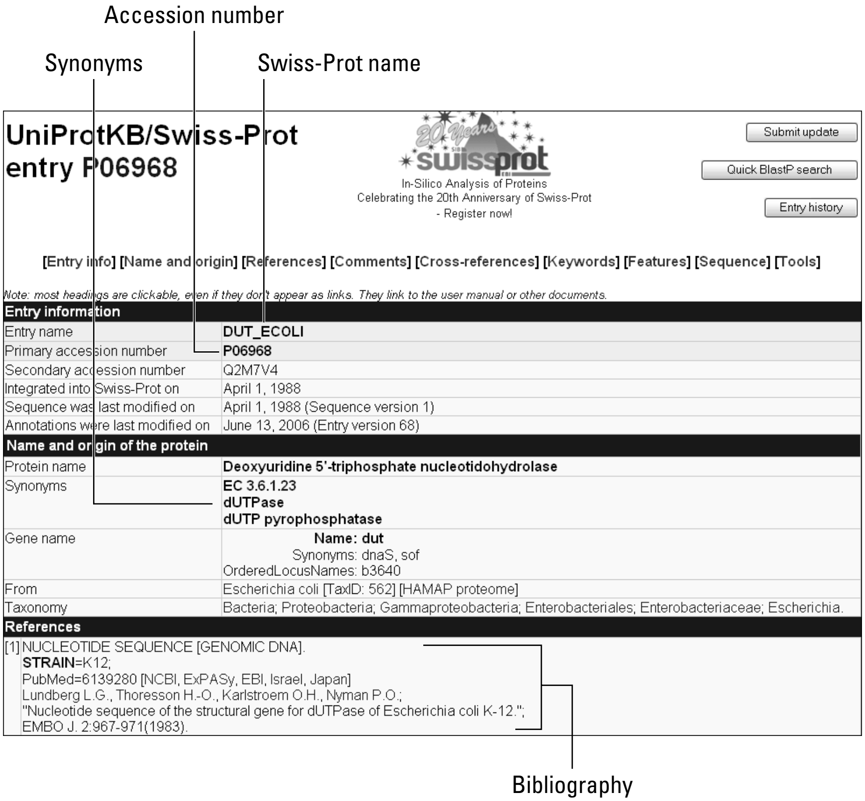
- The top of the entry contains the entry name `DUT_ECOLI` on the first line and the unique identifier `P06968` (called a *primary accession number*) on the second line. As with PubMed identifiers, these codes are worth writing down; they're used to cross-reference related entries in other databases.
- The top section offers a biochemical description of the protein, including its standard name, its international Enzyme Committee number ("E.C." does not mean "E. coli"), as well as a couple of synonyms. These synonyms can be very useful to broaden your search for relevant articles in Medline. Then comes a list of bibliographic references relevant to the sequencing of the protein or some functional studies. Note that the PubMed ID is given for all cited articles to allow for easy cross-referencing.
- The middle section offers a whole series of links to various functional classification schemes — including relevant protein domains, 3-D structures, and functional signatures. (We describe this section in more detail in Chapter 4.)
- The bottom section — the *sequence* section — provides you with the actual amino-acid sequence of the protein. In Figure 2-14, you can see that the *E. coli* dUTPase is made up of 151 amino acids, corresponding to a predicted molecular weight of 16,155 Da. (The sequence shown in Figure 2-14 uses the one-letter code format for describing amino acids. For more on the different formats used, see Chapter 1.)

For further studies, you need this amino-acid sequence in the FASTA format — a much simpler format that doesn't bother with punctuation. See the sidebar "The FASTA (and RAW) format," elsewhere in this chapter, for more information.

3. **To get the FASTA format, click the FASTA Format link, on the extreme right of the very bottom of the entry.**

   The FASTA format for your sequence is shown in Figure 2-15.

In just three easy steps, you've gone all the way from an informal definition such as dUTPase to the complete description of the protein that performs this function in everybody's "favorite" bacterium. (And you thought bioinformatics was complicated?)



UniProtKB/Swiss-Prot entry P06968

In-Silico Analysis of Proteins
Celebrating the 20th Anniversary of Swiss-Prot
- Register now!

[Entry info] [Name and origin] [References] [Comments] [Cross-references] [Keywords] [Features] [Sequence] [Tools]

Note: most headings are clickable, even if they don't appear as links. They link to the user manual or other documents.

Entry information	
Entry name	**DUT_ECOLI**
Primary accession number	**P06968**
Secondary accession number	Q2M7V4
Integrated into Swiss-Prot on	April 1, 1988
Sequence was last modified on	April 1, 1988 (Sequence version 1)
Annotations were last modified on	June 13, 2006 (Entry version 68)
Name and origin of the protein	
Protein name	**Deoxyuridine 5'-triphosphate nucleotidohydrolase**
Synonyms	**EC 3.6.1.23** **dUTPase** **dUTP pyrophosphatase**
Gene name	**Name: dut** Synonyms: dnaS, sof OrderedLocusNames: b3640
From	Escherichia coli [TaxID: 562] [HAMAP proteome]
Taxonomy	Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; Escherichia.
References	

[1] NUCLEOTIDE SEQUENCE [GENOMIC DNA].
STRAIN=K12;
PubMed=6139280 [NCBI, ExPASy, EBI, Israel, Japan]
Lundberg L.G., Thoresson H.-O., Karlstroem O.H., Nyman P.O.;
"Nucleotide sequence of the structural gene for dUTPase of Escherichia coli K-12.";
EMBO J. 2:967-971(1983).

Figure 2-12: Swiss-Prot entry for the E. coli dUTPase protein (upper section).

More advanced ways to retrieve protein sequences

Identifying the right protein in Swiss-Prot isn't always so easy because the information you start with may not be specific enough. For instance, the Search feature we outline in the previous section may not work particularly well if the same term is found in different sections of the Swiss-Prot entry. To get better results, you have to use some field restriction (which we describe in the context of PubMed searches).

This type of restricted search always depends on the organization of each particular database, as well as on the idiosyncrasies of search software. For Swiss-Prot, the steps are as follows:

1. **Point your browser to** `www.expasy.org/sprot/`.

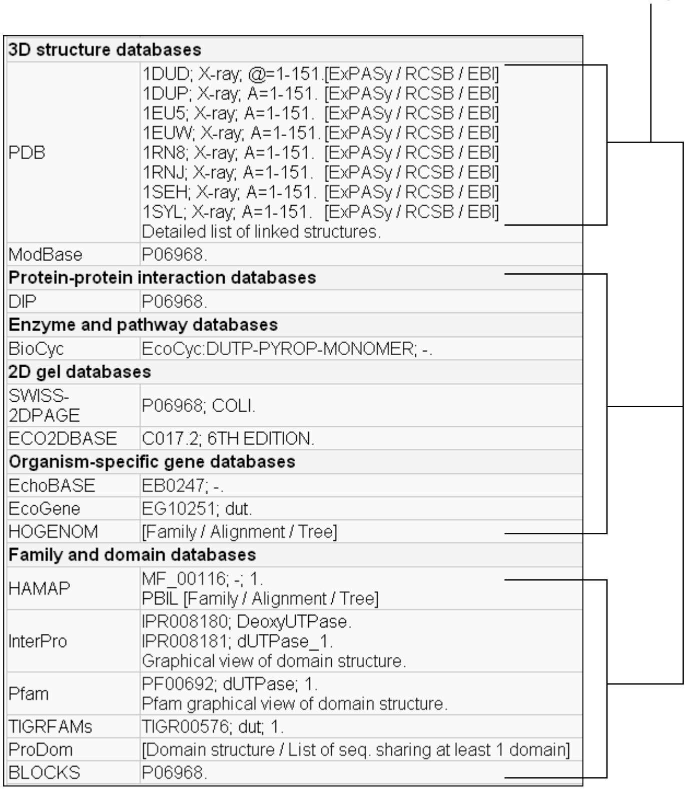
   You'll see the by-now-familiar Swiss-Prot database home page (refer to Figure 2-10).

2. **Click the** Advanced Search in the UniProt Knowledgebase **link (located near the bottom of the opening screen, below the *Access to the UniProt Knowledgebase* banner).**

   The Advanced Search page appears (Figure 2-16). Using this electronic form, you can retrieve protein sequences by using keywords associated with three specific fields: Description, Gene name, and Organism.

3. **For the purposes of our example, type** dUTPase **in the Description field, choose baker's yeast from the Organism drop-down menu, and click the Submit Query button. (Refer to Figure 2-16.)**

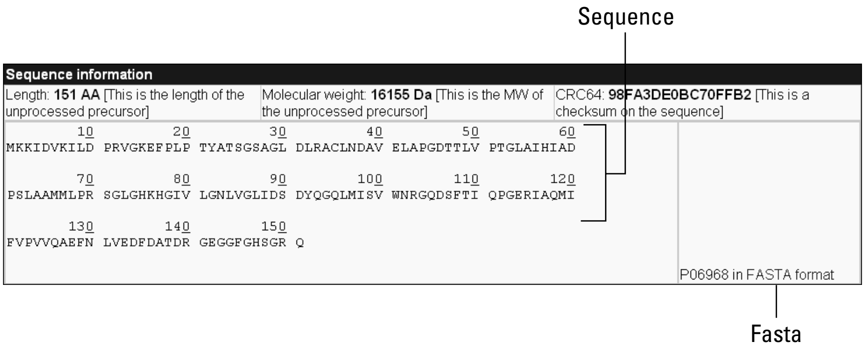
   Your results appear on a new page. If you use the dUTPase example, you'll successfully retrieve the yeast dUTPase protein, no ifs, ands, or buts — that's to say, exactly the protein you were looking for. Its Swiss-Prot entry name is `DUT_YEAST`, and its accession number is `P33317`.



Links to relevant entries in other databases (structure, biochemistry, signatures, . . . , etc.)

3D structure databases	
PDB	1DUD; X-ray; @=1-151.[ExPASy / RCSB / EBI] 1DUP; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EU5; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EUW; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1RN8; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1RNJ; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SEH; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SYL; X-ray; A=1-151. [ExPASy / RCSB / EBI] Detailed list of linked structures.
ModBase	P06968.
Protein-protein interaction databases	
DIP	P06968.
Enzyme and pathway databases	
BioCyc	EcoCyc:DUTP-PYROP-MONOMER; -.
2D gel databases	
SWISS-2DPAGE	P06968; COLI.
ECO2DBASE	C017.2; 6TH EDITION.
Organism-specific gene databases	
EchoBASE	EB0247; -.
EcoGene	EG10251; dut.
HOGENOM	[Family / Alignment / Tree]
Family and domain databases	
HAMAP	MF_00116; -; 1. PBIL [Family / Alignment / Tree]
InterPro	IPR008180; DeoxyUTPase. IPR008181; dUTPase_1. Graphical view of domain structure.
Pfam	PF00692; dUTPase; 1. Pfam graphical view of domain structure.
TIGRFAMs	TIGR00576; dut; 1.
ProDom	[Domain structure / List of seq. sharing at least 1 domain]
BLOCKS	P06968.

Figure 2-13: Swiss-Prot entry for the E. coli dUTPase protein (middle section).



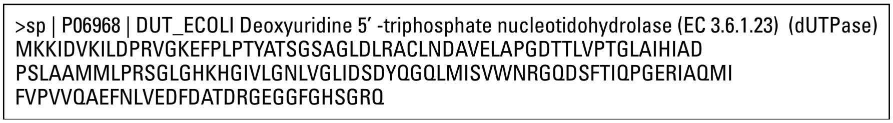
Sequence information

Length: **151 AA** [This is the length of the unprocessed precursor] | Molecular weight: **16155 Da** [This is the MW of the unprocessed precursor] | CRC64: **98FA3DE0BC70FFB2** [This is a checksum on the sequence]

```
        10         20         30         40         50         60
MKKIDVKILD PRVGKEFPLP TYATSGSAGL DLRACLNDAV ELAPGDTTLV PTGLAIHIAD

        70         80         90        100        110        120
PSLAAMMLPR SGLGHKHGIV LGNLVGLIDS DYQGQLMISV WNRGQDSFTI QPGERIAQMI

       130        140        150
FVPVVQAEFN LVEDFDATDR GEGGFGHSGR Q
```

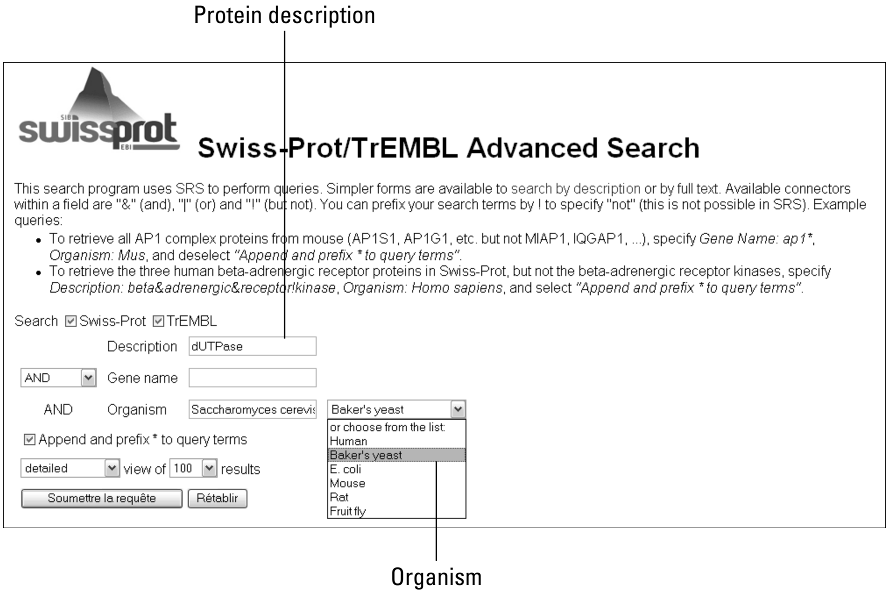
P06968 in FASTA format

Figure 2-14: Amino-acid sequence of the E. coli dUTPase protein (bottom section).



```
>sp | P06968 | DUT_ECOLI Deoxyuridine 5' -triphosphate nucleotidohydrolase (EC 3.6.1.23)  (dUTPase)
MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAPGDTTLVPTGLAIHIAD
PSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWNRGQDSFTIQPGERIAQMI
FVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ
```

Figure 2-15: E. coli dUTPase protein sequence in FASTA format.



Swiss-Prot/TrEMBL Advanced Search

This search program uses SRS to perform queries. Simpler forms are available to search by description or by full text. Available connectors within a field are "&" (and), "|" (or) and "!" (but not). You can prefix your search terms by ! to specify "not" (this is not possible in SRS). Example queries:

- To retrieve all AP1 complex proteins from mouse (AP1S1, AP1G1, etc. but not MIAP1, IQGAP1, ...), specify *Gene Name: ap1**, *Organism: Mus*, and deselect *"Append and prefix * to query terms"*.
- To retrieve the three human beta-adrenergic receptor proteins in Swiss-Prot, but not the beta-adrenergic receptor kinases, specify *Description: beta&adrenergic&receptor!kinase*, *Organism: Homo sapiens*, and select *"Append and prefix * to query terms"*.

Search ☑ Swiss-Prot ☑ TrEMBL

Description: dUTPase

AND Gene name:

AND Organism: Saccharomyces cerevis | Baker's yeast

☑ Append and prefix * to query terms

detailed view of 100 results

Figure 2-16: Using Swiss-Prot Advanced search.

The FASTA (and RAW) format

FASTA is the name of a popular sequence-alignment-and-database-scanning program created by W.R. Pearson and D.J. Lipman in 1988 (you can use your brand new PubMed skills to find the original article). The sequences used by FASTA have to obey the following format:

```
>My_Sequence_Name

ARCGTCRGCKINTANDRGCKINTAND

CKINTANDARCGTCRGCKINTANDRG

CKINTAND
```

The line starting with > (the *definition line*) contains a unique identifier followed by an optional short definition. The lines that follow it contain the DNA or protein sequence (in one-letter code) until the next > character in the file indicates the beginning of a new sequence.

Because FASTA is easy to parse, this format has become hugely popular — and is now the default input format for much sequence analysis software, including BLAST and CLUSTALW.

Be aware, though, that programs using FASTA-formatted sequences as input are sometimes case-sensitive. Here are some pointers:

- Always use CAPITAL letters for the one-letter codes.
- When using FASTA-formatted sequences on a PC, always use the TEXT option of your preferred word-processing software (that is, skip the formatting and use nothing but ASCII characters).
- When displaying these sequences as a word-processing document, use the Courier font for easy alignment.

Some programs analyzing one sequence at a time work with the *RAW* format. This is simply the sequence part of the FASTA format, without the definition line — but machines can be finicky. Using the FASTA format when the RAW format is required may cause an error — or some of the definition line may end up included in the protein or DNA sequence (!).

Retrieving a list of related protein sequences

Many questions in molecular biology (your dissertation topic, your own research, or your personal interests) require downloading a large collection of similar protein sequences, all related to the same function, rather than just one sequence. These biological questions typically include the detection of conserved functional motifs (segments of sequences that look the same in proteins with the same function), the simultaneous alignment of multiple sequences, the assessment of their variability, or *phylogenetic* studies — how sequences relate to each other through evolution.

Building your own specialized sequence data file, right on your PC is easy and enables you to use it with other analysis programs accessible from other bioinformatic sites. To continue with our previous example, we now show you how to build a file containing all well-documented dUTPase protein sequences. To make it usable by other programs, you generate this file in FASTA format. Here's the procedure:

First, go back to the Advanced Search page of the ExPASy server:

1. **Point your browser to** `www.expasy.org/sprot/` **and click the** Advanced Search in the UniProt Knowledgebase **link.**
2. **In the Search line — directly above the Description window — keep the Swiss-Prot box checked but deselect the TrEMBL box.**

   TrEMBL is a database made up of unsupervised computer translations of new DNA sequences, Swiss-Prot only includes entries validated by expert curators. Restricting the search to Swiss-Prot thus ensures — to the best of our knowledge — that all returned proteins are *actual* dUTPases.
3. **Type** dUTPase **in the Description window.**

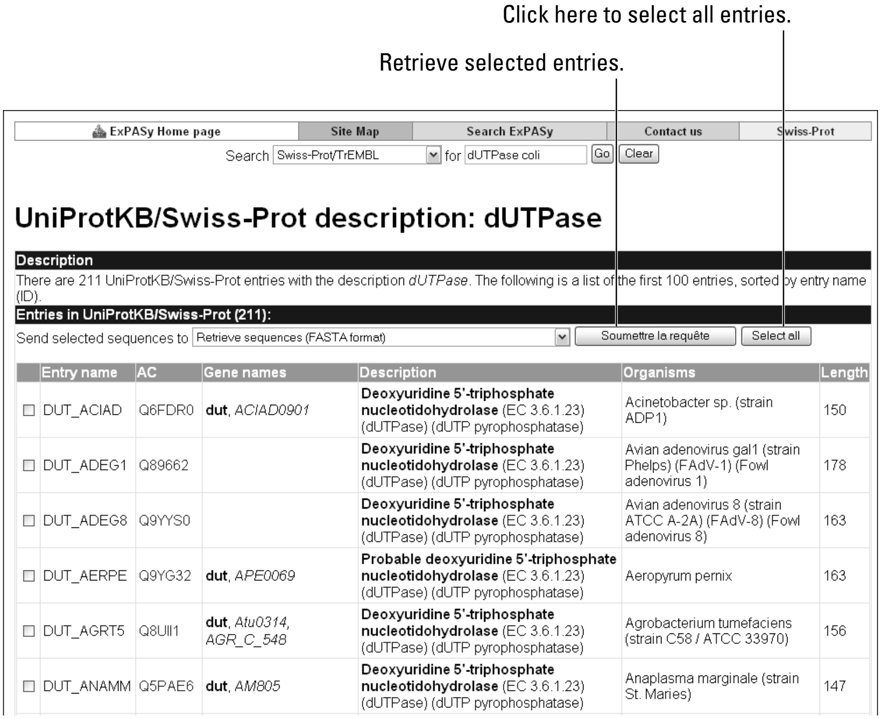
   Be sure that you don't type in anything else. In particular, don't put anything in the Organism window.
4. **Click the Submit Query button.**

   A Results page appears, as shown in Figure 2-17. This particular search (at the time of writing) yielded 211 Swiss-Prot entries. (You may get more entries when you get around to doing your own search.) Scrolling down the list, you can see that all the entries look like fine and upstanding dUTPase protein sequences. At this point, you can click each of the individual names (such as `DUT_ADEG1`) to access a complete ID card. (For a sense of what you can find on such an ID card, refer to Figures 2-12, 2-13 and 2-14.)

But to complete your assignment on dUTPases, for instance, you now have to gather all these sequences into a single file. This is easy; just do the following:

1. **Perform Steps 1 through 4 in the preceding list, and then click the Select All button at the top-right of the Results screen (refer to Figure 2-17) or check the boxes for some specific sequences you are interested in.**
2. **Don't waste any time here and don't change anything; simply click the Send query button (that's *soumettre la requête* in French).**

   The Download page appears, as shown in Figure 2-18. Your sequences are ready to be saved to your hard drive.



Click here to select all entries.

Retrieve selected entries.

Search Swiss-Prot/TrEMBL for dUTPase coli Go Clear

UniProtKB/Swiss-Prot description: dUTPase

Description

There are 211 UniProtKB/Swiss-Prot entries with the description *dUTPase*. The following is a list of the first 100 entries, sorted by entry name (ID).

Entries in UniProtKB/Swiss-Prot (211):

Send selected sequences to Retrieve sequences (FASTA format) Soumettre la requête Select all

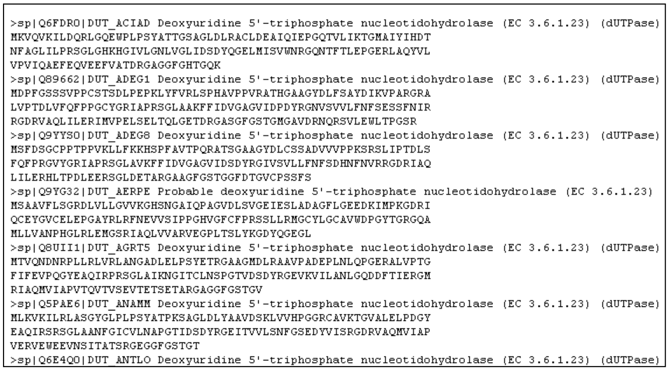
	Entry name	AC	Gene names	Description	Organisms	Length
□	DUT_ACIAD	Q6FDR0	**dut**, *ACIAD0901*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Acinetobacter sp. (strain ADP1)	150
□	DUT_ADEG1	Q89662		**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Avian adenovirus gal1 (strain Phelps) (FAdV-1) (Fowl adenovirus 1)	178
□	DUT_ADEG8	Q9YYS0		**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Avian adenovirus 8 (strain ATCC A-2A) (FAdV-8) (Fowl adenovirus 8)	163
□	DUT_AERPE	Q9YG32	**dut**, *APE0069*	**Probable deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Aeropyrum pernix	163
□	DUT_AGRT5	Q8UII1	**dut**, *Atu0314*, *AGR_C_548*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Agrobacterium tumefaciens (strain C58 / ATCC 33970)	156
□	DUT_ANAMM	Q5PAE6	**dut**, *AM805*	**Deoxyuridine 5'-triphosphate nucleotidohydrolase** (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase)	Anaplasma marginale (strain St. Maries)	147

Figure 2-17: Top of the list of all dUTPase proteins found in Swiss-Prot.

You can save the content of this page in various ways, depending on the Internet browser version you're using. We recommend the following procedure because it always works — and it doesn't add mysterious extra characters to the resulting file!

1. **Choose the Select All feature from your browser menu (in Internet Explorer, choose Edit➪Select All) and then choose Copy. (Again, in Internet Explorer you choose Edit➪Copy.)**
2. **Open a new document in your word-processing program (Microsoft Word, for example).**
3. **Choose Paste from the Edit menu of your word-processing program.**
4. **Reformat the whole document with a Courier font (8 or 10 point) to realign the sequences.**
5. **Finally, save your document as** `dUTPaseDB.txt`**, using the Save As type option *text only*** (`*.txt`).

You now have, on your own personal computer, all the well-characterized dUTPase protein sequences known to man. Keep this file on your hard drive; you need it later in this chapter.



```
>sp|Q6FDR0|DUT_ACIAD Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MKVQVKILDQRLGQEWPLPSYATTGSAGLDLRACLDEAIQIEPGQTVLIKTGMAIYIHDT
NFAGLILPRSGLGHKHGIVLGNLVGLIDSDYQGELMISVWNRGQNTFTLEPGERLAQYVL
VPVIQAEFEQVEEFVATDRGAGGFGHTGQK
>sp|Q89662|DUT_ADEG1 Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA
LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR
RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR
>sp|Q9YYS0|DUT_ADEG8 Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MSFDSGCPPTPPVKLLFKKHSPFAVTPQRATSGAAGYDLCSSADVVVPPKSRSLIPTDLS
FQFPRGVYGRIAPRSGLAVKFFIDVGAGVIDSDYRGIVSVLLFNFSDHNFNVRRGDRIAQ
LILERHLTPDLEERSGLDETARGAAGFGSTGGFDTGVCPSSFS
>sp|Q9YG32|DUT_AERPE Probable deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23)
MSAAVFLSGRDLVLLGVVKGHSNGAIQPAGVDLSVGEIESLADAGFLGEEDKIMPKGDRI
QCEYGVCELEPGAYRLRFNEVVSIPPGHVGFCFPRSSLLRMGCYLGCAVWDPGYTGRGQA
MLLVANPHGLRLEMGSRIAQLVVARVEGPLTSLYKGDYQGEGL
>sp|Q8UII1|DUT_AGRT5 Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MTVQNDNRPLLRLVRLANGADLELPSYETRGAAGMDLRAAVPADEPLNLQPGERALVPTG
FIFEVPQGYEAQIRPRSGLAIKNGITCLNSPGTVDSDYRGEVKVILANLGQDDFTIERGM
RIAQMVIAPVTQVTVSEVTETSETARGAGGFGSTGV
>sp|Q5PAE6|DUT_ANAMM Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
MLKVKILRLASGYGLPLPSYATPKSAGLDLYAAVDSKLVVHPGGRCAVKTGVALELPDGY
EAQIRSRSGLAANFGICVLNAPGTIDSDYRGEITVVLSNFGSEDYVISRGDRVAQMVIAP
VERVEWEEVNSITATSRGEGGFGSTGT
>sp|Q6E4Q0|DUT_ANTLO Deoxyuridine 5'-triphosphate nucleotidohydrolase (EC 3.6.1.23) (dUTPase)
```

Figure 2-18: FASTA sequences ready to be saved to your hard drive.

Retrieving DNA Sequences

Protein sequences are simple objects with a relatively narrow range of sizes (they're about 300 amino acids long, plus or minus 200, except for a few giant ones), clearly defined boundaries, and specific functional attributes. Furthermore, proteins of microbes or higher *eukaryotes* (animal and plants) have roughly the same properties.

As you might expect, the corresponding gene (DNA) sequences get more varied and complex in higher animals. Gene sizes in humans may vary from microbe-like lengths (a few thousand bp) to several hundred thousand bp.

Not all DNA is coding for protein

Various types of DNA sequences are involved in defining a gene:

- Regulatory regions (usually preceding the coding region)
- Untranslated regions that precede and follow the coding regions
- The protein-coding region

In eukaryotes (yeast, plants, animals), the protein-coding region is divided into a variable number of *exons* — gene segments that contribute to the final protein — interspersed with *introns* — gene segments that do not.

As a consequence, working with DNA sequences is always trickier than working with protein sequences. Go to Chapters 3 and 5 to find out more about the architecture of genes.

Beware of parasite characters in downloaded sequence files

In the dUTPAse example, we didn't use the File➪Save As command on the Internet browser main menu to download the content of the window because files saved using the File➪Save As command have some problems:

- They may contain some hidden *parasite characters* — Control/Alt/Shift + something — often displayed as **<PRE>** at the beginning of the file — which corrupts the FASTA format.
- They're trickier to reopen by your PC word-processing software, because they don't have the right extension (for instance `.doc`), or ask you to choose an obscure encoding scheme (about which we know nothing, so we can't advise you) to load the file.

It is our experience that using a browser's File➪Save As option produces unpredictable results, depending on the browser type, version, or implementation. In general, it's a good and wise practice to inspect the sequence data files that you download from the Internet for unexpected leading or trailing signs. For most sequence-analysis programs, FASTA-formatted sequence files must begin with a definition line, such as

```
>P0343456
     My_Sequence_definition
```

and nothing else! Any leading character (even blank ones) that differs from > may produce an error. (For instance, the definition line might be considered part of the protein sequence.) Sequence-data files also have to end with a final `<New Line>` character (showing as a blank line).

The good news is that usually no constraints restrict the length of the definition and sequence lines (but use reasonable numbers, no more than about 60 to 100 characters long, to be safe). Except for the constraint of having > as the first character of any definition line, you can freely use blank characters, such as `<Space>`, `<Tab>`, and line delimiters without interfering with the parsing of the sequences.

Finally, do not use characters that are not among the standard amino-acid codes such as `<->` or `<*>`. They aren't treated in a consistent way by different analysis programs. They're skipped (deleting a position), replaced by `X`, or may simply cause an error. (For more on standard amino-acid codes, see Chapter 1.)

Going from protein sequences to DNA sequences

In databases, the correspondence between protein and DNA sequences is *not* one-to-one. Many different — even non-overlapping — DNA sequences can be linked to the same protein or gene name, as the following list makes clear (flip ahead to Figures 3-1 and 3-2):

- The *primary transcript* — that is generated by copying the DNA sequence of a gene from beginning to end, (including exons + introns)
- The *mature transcript* — the mRNA, generated from the primary transcript by discarding the introns)
- The strict protein-coding region — the open reading frame or ORF
- Numerous types of partial sequences

Life is a lot simpler if you can stick to working with just protein sequences, but we realize that situations will arise that will require you to go back to the DNA sequence. The most common situation of this kind is when you want to amplify a gene to transfer it to another organism — using a technique called Polymerase Chain Reaction or PCR — and then either synthesize the corresponding protein or induce specific mutations. Your problem, succinctly stated, is this: *Given a protein sequence, how can I retrieve the DNA sequence encompassing its coding region?*

Retrieving the DNA sequence relevant to my protein

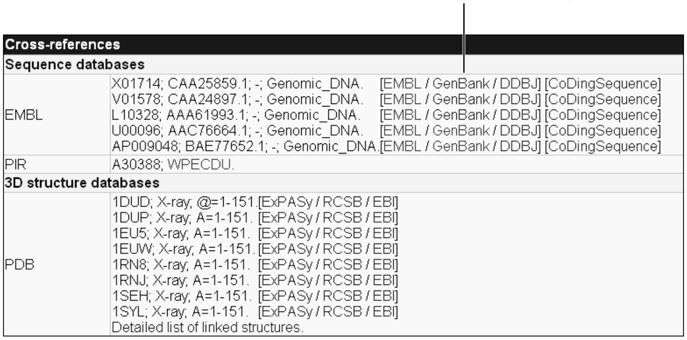
Imagine that you want to retrieve enough DNA sequence to clone the dUTPase gene of *E. coli.* Sounds tricky, right? We show you how it's done in the following steps:

1. **Point your browser to** `www.expasy.org/sprot/`.
2. **To access the *E. coli* dUTPase entry quickly, simply enter the accession number (**P06968**) in the Search window at the top of the page and then click the Search button.**

   Again, we have the information page devoted to protein P06968 at hand (refer to Figure 2-12.).
3. **Click the** Cross-References **link near the top of the form.**

   Your browser jumps to the relevant section, as shown in Figure 2-19. The Cross-References section consists of successive groups of lines introduced by keywords such as `EMBL`, `PIR`, `PDB`, and so on. All these keywords correspond to databases of various kinds that can give you additional information about your protein sequence. By clicking the different underlined words, you jump right to the corresponding entry in those different databases. This web of links between different sites that concern the same object is, in fact, a web of cross-references. Take the EMBL group of lines, for instance: There are four of them, starting with obscure numbers and followed by a set of four underlined links to EMBL, GenBank, DDBJ (those are databases), and CoDingSequence (the precise DNA segment coding for the protein). Let's see what happens when we use them.

Click here to get the DNA sequence.



Cross-references	
Sequence databases	
EMBL	X01714; CAA25859.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] V01578; CAA24897.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] L10328; AAA61993.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] U00096; AAC76664.1; -; Genomic_DNA. [EMBL / GenBank / DDBJ] [CoDingSequence] AP009048; BAE77652.1; -; Genomic_DNA.[EMBL / GenBank / DDBJ] [CoDingSequence]
PIR	A30388; WPECDU.
3D structure databases	
PDB	1DUD; X-ray; @=1-151.[ExPASy / RCSB / EBI] 1DUP; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EU5; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1EUW; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1RN8; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1RNJ; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SEH; X-ray; A=1-151. [ExPASy / RCSB / EBI] 1SYL; X-ray; A=1-151. [ExPASy / RCSB / EBI] Detailed list of linked structures.

Figure 2-19: Cross-References section of the P06968 information page.

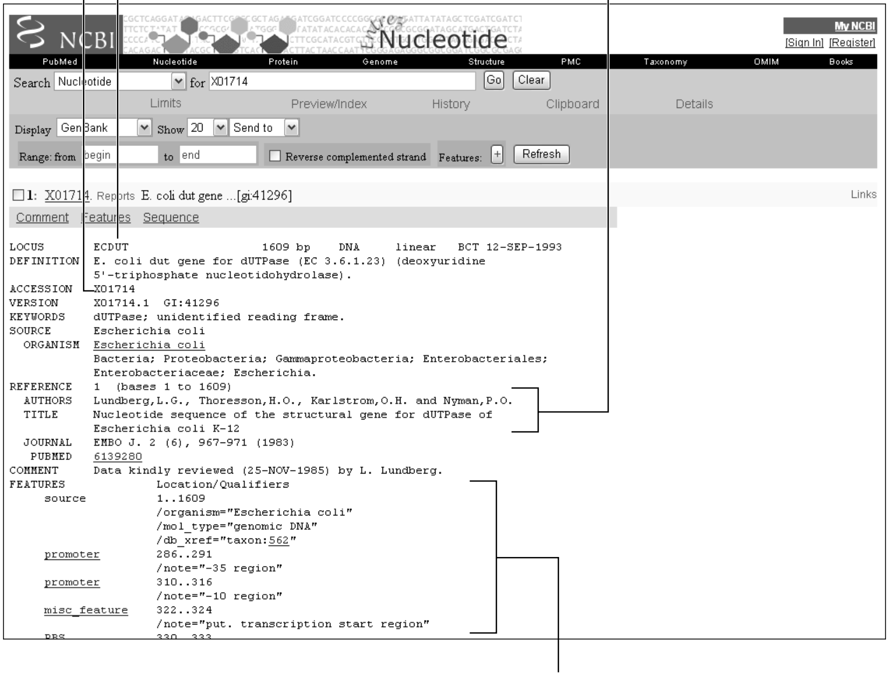
4. **Click the GenBank link in the EMBL row's first line.**

   The GenBank record (Accession # X01714) corresponding to protein P06968 appears, as shown in Figure 2-20. It is several pages long.

Accession number

GenBank name

Bibliography



□1: X01714. Reports E. coli dut gene ...[gi:41296]

```
LOCUS       ECDUT                   1609 bp    DNA     linear   BCT 12-SEP-1993
DEFINITION  E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
            5'-triphosphate nucleotidohydrolase).
ACCESSION   X01714
VERSION     X01714.1  GI:41296
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
```

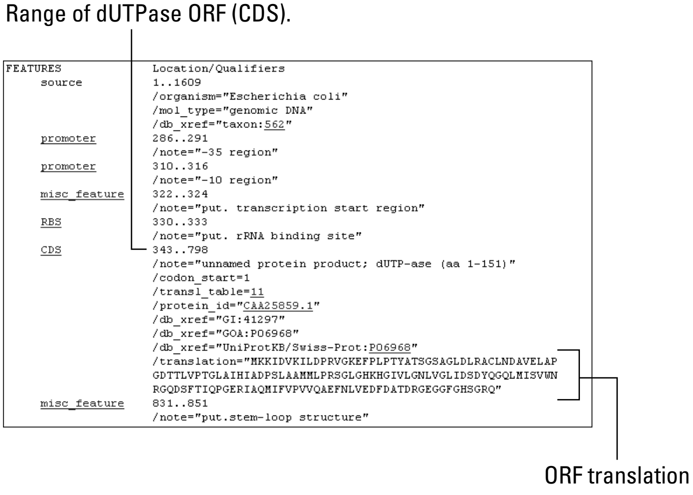
Features

Figure 2-20: Top of GenBank entry ECDUT/ X01714.

Roughly speaking, a GenBank entry consists of four parts. Here's how they look for our dUTPase example:

- The *locus name* (`ECDUT`) — an arbitrary identifier — is followed by a short definition line and a unique accession number (`X01714`). This number is the most stable identifier for the entry; we recommend that you write it down because it's used to cross-reference related entries in other databases (which is what we did with Swiss-Prot). A few lines below, the `SOURCE` and `ORGANISM` fields describe the biological origin of the DNA sequence.
- The *Reference section* lists article(s) relevant to the sequence determination. This list can be quite long for large sequences.
- The *Features section* lists the definitions and exact ranges of multiple types of elements that have been recognized in the sequence. Our example — the X01714 entry — includes promoter elements, ribosome binding sites (RBS), and protein coding segments (CDS). Figure 2-21 shows — right next to the keyword CDS — the limits of the dUTPase ORF as `343..798`. The CDS range is followed by the name and the amino-acid translation of the corresponding protein.
- The *Sequence section* rounds out the GenBank entry, where the nucleotides are listed between the Origin keyword and the final `//` that signals the very end of the entry. Numbering is provided to help relate the location of the dUTPase ORF (343-798) to the actual nucleotide sequence. (See Figure 2-22.)

When you've read through the entry, you can save the whole thing to your hard drive by using the steps we defined in the preceding list for saving Swiss-Prot sequences.

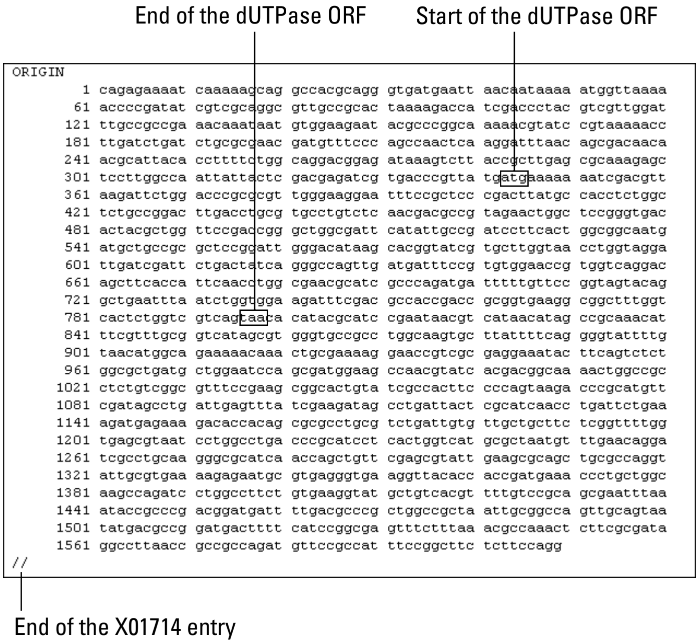


Range of dUTPase ORF (CDS).

```
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="unnamed protein product; dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="GOA:P06968"
                     /db_xref="UniProtKB/Swiss-Prot:P06968"
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc_feature    831..851
                     /note="put.stem-loop structure"
```

ORF translation

Figure 2-21: The Features section of GenBank entry ECDUT/ X01714.



```
ORIGIN
        1 cagagaaaat caaaaagcag gccacgcagg gtgatgaatt aacaataaaa atggttaaaa
       61 accccgatat cgtcgcaggc gttgccgcac taaaagacca tcgaccctac gtcgttggat
      121 ttgccgccga aacaaataat gtggaagaat acgcccggca aaaacgtatc cgtaaaaacc
      181 ttgatctgat ctgcgcgaac gatgtttccc agccaactca aggatttaac agcgacaaca
      241 acgcattaca ccttttctgg caggacggag ataaagtctt accgcttgag cgcaaagagc
      301 tccttggcca attattactc gacgagatcg tgacccgtta tgatgaaaaa aatcgacgtt
      361 aagattctgg acccgcgcgt tgggaaggaa tttccgctcc cgacttatgc cacctctggc
      421 tctgccggac ttgacctgcg tgcctgtctc aacgacgccg tagaactggc tccgggtgac
      481 actacgctgg ttccgaccgg gctggcgatt catattgccg atccttcact ggcggcaatg
      541 atgctgccgc gctccggatt gggacataag cacggtatcg tgcttggtaa cctggtagga
      601 ttgatcgatt ctgactatca gggccagttg atgatttccg tgtggaaccg tggtcaggac
      661 agcttcacca ttcaacctgg cgaacgcatc gcccagatga tttttgttcc ggtagtacag
      721 gctgaattta atctggtgga agatttcgac gccaccgacc gcggtgaagg cggctttggt
      781 cactctggtc gtcagtaaca catacgcatc cgaataacgt cataacatag ccgcaaacat
      841 ttcgtttgcg gtcatagcgt gggtgccgcc tggcaagtgc ttattttcag gggtattttg
      901 taacatggca gaaaaacaaa ctgcgaaaag gaaccgtcgc gaggaaatac ttcagtctct
      961 ggcgctgatg ctggaatcca gcgatggaag ccaacgtatc acgacggcaa aactggccgc
     1021 ctctgtcggc gtttccgaag cggcactgta tcgccacttc cccagtaaga cccgcatgtt
     1081 cgatagcctg attgagttta tcgaagatag cctgattact cgcatcaacc tgattctgaa
     1141 agatgagaaa gacaccacag cgcgcctgcg tctgattgtg ttgctgcttc tcggttttgg
     1201 tgagcgtaat cctggcctga cccgcatcct cactggtcat gcgctaatgt ttgaacagga
     1261 tcgcctgcaa gggcgcatca accagctgtt cgagcgtatt gaagcgcagc tgcgccaggt
     1321 attgcgtgaa aagagaatgc gtgagggtga aggttacacc accgatgaaa ccctgctggc
     1381 aagccagatc ctggccttct gtgaaggtat gctgtcacgt tttgtccgca gcgaatttaa
     1441 ataccgcccg acggatgatt ttgacgcccg ctggccgcta attgcggcca gttgcagtaa
     1501 tatgacgccg gatgactttt catccggcga gtttctttaa acgccaaact cttcgcgata
     1561 ggccttaacc gccgccagat gttccgccat ttccggcttc tcttccagg
//
```

Figure 2-22: The Sequence section of GenBank entry ECDUT/X01714.

However, to use the nucleotide sequence as input for other programs (for instance, to design primers), you may need to isolate it from the text part of the entry and convert it in FASTA format. Here's how that's done:

1. **Scroll back to the top of the page for the ECDUT/X01714 entry.**

   Refer to Figure 2-20 for what your screen should look like.

2. **Choose FASTA from the Display drop-down menu, as shown in Figure 2-23.**

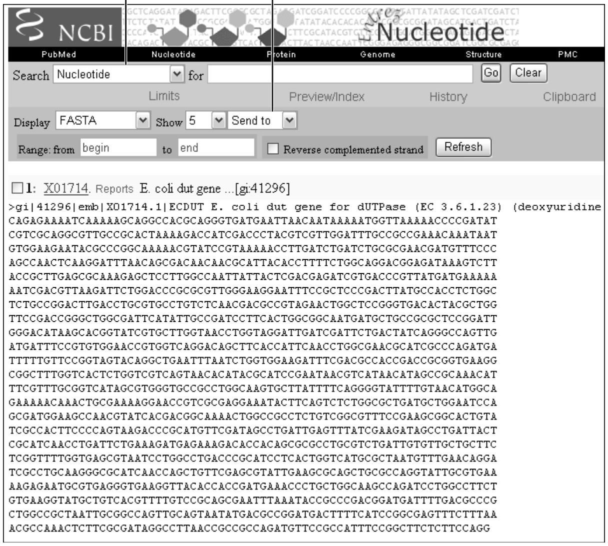
3. **Transform the content of this window into plain text by choosing Text from the drop-down menu located on the far right of the menu bar.**

4. **Save the FASTA sequence by using the following protocol:**

   a. In the Edit menu of your Web browser, click Select All and then click Copy.

   b. Open a default Word document and, in the Edit menu of Word, click Paste. Then select a Courier font (8 or 10).

   c. Finally, save your document as dUTPaseDNA.txt by choosing the Save as type option text only (`*.txt`).



Send to TEXT for printer-friendly display.

Select FASTA format.

Search Nucleotide for

Display FASTA Show 5 Send to

Range: from begin to end □ Reverse complemented strand Refresh

□ 1: X01714. Reports E. coli dut gene ...[gi:41296]

```
>gi|41296|emb|X01714.1|ECDUT E. coli dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
CAGAGAAAATCAAAAAGCAGGCCACGCAGGGTGATGAATTAACAATAAAAATGGTTAAAAACCCCGATAT
CGTCGCAGGCGTTGCCGCACTAAAAGACCATCGACCCTACGTCGTTGGATTTGCCGCCGAAACAAATAAT
GTGGAAGAATACGCCCGGCAAAAACGTATCCGTAAAAACCTTGATCTGATCTGCGCGAACGATGTTTCCC
AGCCAACTCAAGGATTTAACAGCGACAACAACGCATTACACCTTTTCTGGCAGGACGGAGATAAAGTCTT
ACCGCTTGAGCGCAAAGAGCTCCTTGGCCAATTATTACTCGACGAGATCGTGACCCGTTATGATGAAAAA
AATCGACGTTAAGATTCTGGACCCGCGCGTTGGGAAGGAATTTCCGCTCCCGACTTATGCCACCTCTGGC
TCTGCCGGACTTGACCTGCGTGCCTGTCTCAACGACGCCGTAGAACTGGCTCCGGGTGACACTACGCTGG
TTCCGACCGGGCTGGCGATTCATATTGCCGATCCTTCACTGGCGGCAATGATGCTGCCGCGCTCCGGATT
GGGACATAAGCACGGTATCGTGCTTGGTAACCTGGTAGGATTGATCGATTCTGACTATCAGGGCCAGTTG
ATGATTTCCGTGTGGAACCGTGGTCAGGACAGCTTCACCATTCAACCTGGCGAACGCATCGCCCAGATGA
TTTTTGTTCCGGTAGTACAGGCTGAATTTAATCTGGTGGAAGATTTCGACGCCACCGACCGCGGTGAAGG
CGGCTTTGGTCACTCTGGTCGTCAGTAACACATACGCATCCGAATAACGTCATAACATAGCCGCAAACAT
TTCGTTTGCGGTCATAGCGTGGGTGCCGCCTGGCAAGTGCTTATTTTCAGGGGTATTTTGTAACATGGCA
GAAAAACAAACTGCGAAAAGGAACCGTCGCGAGGAAATACTTCAGTCTCTGGCGCTGATGCTGGAATCCA
GCGATGGAAGCCAACGTATCACGACGGCAAAACTGGCCGCCTCTGTCGGCGTTTCCGAAGCGGCACTGTA
TCGCCACTTCCCCAGTAAGACCCGCATGTTCGATAGCCTGATTGAGTTTATCGAAGATAGCCTGATTACT
CGCATCAACCTGATTCTGAAAGATGAGAAAGACACCACAGCGCGCCTGCGTCTGATTGTGTTGCTGCTTC
TCGGTTTTGGTGAGCGTAATCCTGGCCTGACCCGCATCCTCACTGGTCATGCGCTAATGTTTGAACAGGA
TCGCCTGCAAGGGCGCATCAACCAGCTGTTCGAGCGTATTGAAGCGCAGCTGCGCCAGGTATTGCGTGAA
AAGAGAATGCGTGAGGGTGAAGGTTACACCACCGATGAAACCCTGCTGGCAAGCCAGATCCTGGCCTTCT
GTGAAGGTATGCTGTCACGTTTTGTCCGCAGCGAATTTAAATACCGCCCGACGGATGATTTTGACGCCCG
CTGGCCGCTAATTGCGGCCAGTTGCAGTAATATGACGCCGGATGACTTTTCATCCGGCGAGTTTCTTTAA
ACGCCAAACTCTTCGCGATAGGCCTTAACCGCCGCCAGATGTTCCGCCATTTCCGGCTTCTCTTCCAGG
```

Figure 2-23: Nucleotide sequence of GenBank entry ECDUT/X01714 in FASTA format.

Using BLAST to Compare My Protein Sequence to Other Protein Sequences

After you know the basics of retrieving a protein sequence from a database (see the earlier, aptly named "Retrieving Protein Sequences" section), you're ready to perform your first analysis with it. For most people, the next step is to perform a BLAST search.

BLAST (short for *B*asic *L*ocal *A*lignment *S*earch *T*ool — with a name like that, no wonder they shortened it to BLAST) is a great sequence-comparison tool that quickly tells you which of the other known proteins out there has a sequence similar to yours. You can then use this information for a variety of purposes — including the prediction of protein function, 3-D structure and domain organization, or the identification of homologues (similar proteins) in other organisms. (In Chapter 7, we show you how to use BLAST in great detail. At this point, we only want to get you started and show you how easy it is to use.)

BLAST comes to us under the auspices of the National Center for Biotechnology Information (NCBI), already familiar to us as the hosts for the PubMed bibliographic database we cover at the beginning of this chapter. You'll soon discover that using BLAST is just as easy as using PubMed. To start off, follow these steps:

1. **Point your favorite Internet browser to**

   `www.ncbi.nlm.nih.gov/BLAST/`

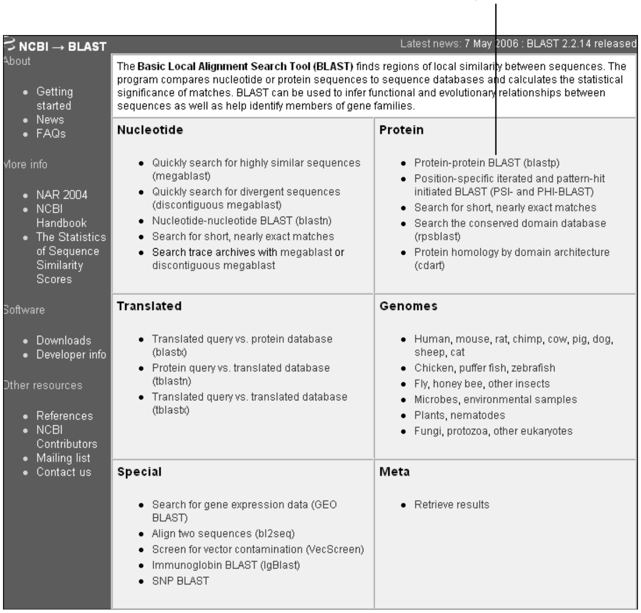
   The BLAST home page — probably the most frequented bioinformatic Web page in the world — appears, as shown in Figure 2-24. Because this is your first time here, keeping things simple is best.

2. **Click the** Protein-Protein BLAST (blastp) **link in the top right.**

   A Query screen appears, as shown in Figure 2-25. At this point, you need a FASTA-formatted protein sequence.

3. **Open the file that contains your dUTPase FASTA-formatted protein sequence.**

   This is the file that you (hopefully) created on your PC by using the steps shown earlier in the "Retrieving a list of related protein sequences" section of this chapter.



Standard protein BLAST

The **Basic Local Alignment Search Tool (BLAST)** finds regions of local similarity between sequences. The program compares nucleotide or protein sequences to sequence databases and calculates the statistical significance of matches. BLAST can be used to infer functional and evolutionary relationships between sequences as well as help identify members of gene families.

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

Special

- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST

Meta

- Retrieve results

Figure 2-24: NCBI BLAST entry page.

4. **Using your browser's Edit menu, copy and paste ONE of the protein sequences (with its definition line) into the BLAST Search window.**

   We used `DUT_ADEG1`. If you don't have the protein sequence ready, open a new navigation window and go get this sequence on the ExPASy server. Here's the quick detour:

   a. Point your browser to `www.expasy.org/sprot/`.

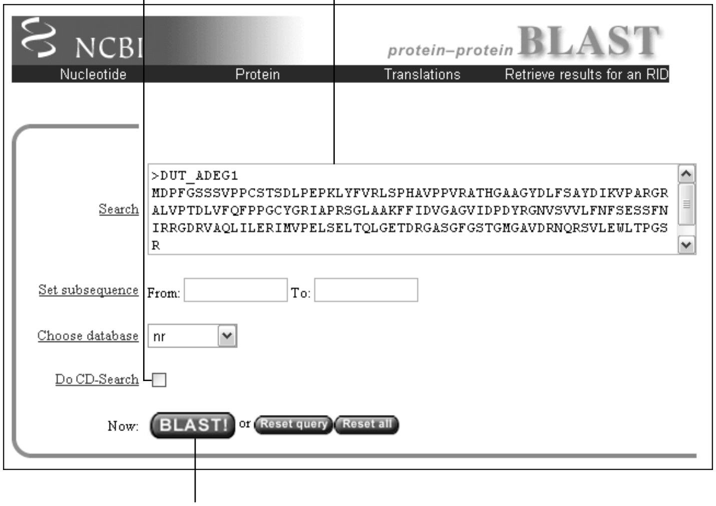
   b. Enter **DUT_ADEG1** in the Search window at the top, and then click Search.

   d. Copy the protein sequence with its header line.

   e. Close that navigator window, and paste the sequence in the BLAST Search window.

Paste your query sequence in here.

Uncheck the CD search box.


NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID

Search

```
>DUT_ADEG1
MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGR
ALVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFN
IRRGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGS
R
```

Set subsequence From: To:

Choose database nr

Do CD-Search □

Now: BLAST! or Reset query Reset all

Click here to run the search.

Figure 2-25: NCBI BLAST query form.

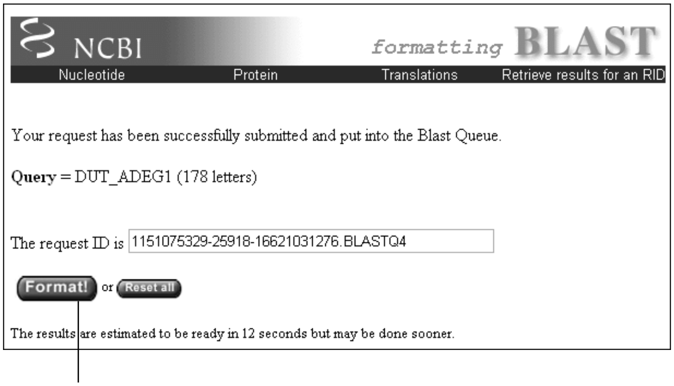
5. **Deselect the Do CD-Search box, but don't change the Choose Database setting.**

   The ***nr*** (for *nonredundant*) default database includes all protein sequences known to man — so you don't have to bother about selecting a specific organism or subset.

6. **Click the BLAST! button.**

   You just launched your first BLAST search. Welcome to the club!

The next form that appears (titled "Formatting BLAST") allows you to change a number of output features. (See Figure 2-26.) For now, keep the defaults as they are.



Your request has been successfully submitted and put into the Blast Queue.

Query = DUT_ADEG1 (178 letters)

The request ID is 1151075329-25918-16621031276.BLASTQ4

The results are estimated to be ready in 12 seconds but may be done sooner.

Click here once and wait.

Figure 2-26: NCBI BLAST "format-while-you-wait" form.

7. **Click the Format! button.**

   BLAST now gives you an estimated running time for your query. If the time of the day is right, this shouldn't take more than a few seconds.

8. **When the Results page appears, scroll down the page until you reach a long list of sequences (Figure 2-27).**

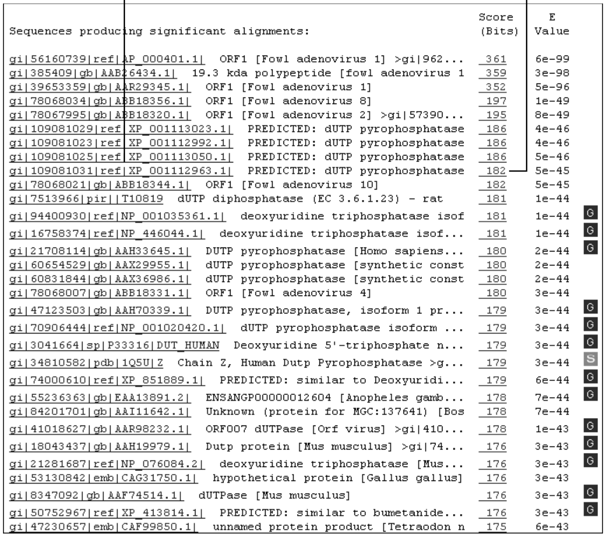
   What you have here are all the sequences that have a significant similarity with your query sequence, ranked by decreasing score values. If you used `DUT_ADEG1` or another protein sequence taken from a database, the best matching protein is probably the one you started with. (This is what happened in Figure 2-27, with the *Fowl adenovirus* dUTPase.) Here the highest similarity score is 361. The score value depends on the length of the most-similar segments that occur between two sequences — as well as on the quality of the match. It is not normalized to 100 percent. (See Chapter 7 to find out more about the significance of BLAST scores.)

   In this list of sequences, you would simply click an underlined identifier to link to the corresponding *nr* (nonredundant) database entry.

9. **Click one of the underlined scores in the second column from the right.**

   This brings you further down in the page (Figure 2-28), to see the alignment between your query sequence and the matching nr sequence of the protein that corresponds to this score.

In our example, the best match is (not surprisingly) an identical one. By going down the list, you can see less-than-perfect matches, slowly degrading as the corresponding score decreases and the E-value increases. The *E-value* is an assessment of the statistical significance of the score. E-values close to 1 are a warning that the conclusion you might draw from the alignments is not reliable. (Go to Chapter 7 to find out more about these issues.)

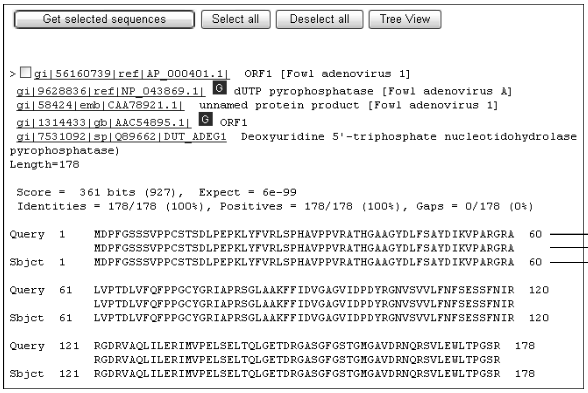


Click here to see the nr entry.

Click here to see the corresponding alignment.

```
                                                                  Score    E
Sequences producing significant alignments:                       (Bits)  Value

gi|56160739|ref|AP_000401.1|  ORF1 [Fowl adenovirus 1] >gi|962...   361   6e-99
gi|385409|gb|AAB26434.1|  19.3 kda polypeptide [fowl adenovirus 1   359   3e-98
gi|39653359|gb|AAR29345.1|  ORF1 [Fowl adenovirus 1]                352   5e-96
gi|78068034|gb|ABB18356.1|  ORF1 [Fowl adenovirus 8]                197   1e-49
gi|78067995|gb|ABB18320.1|  ORF1 [Fowl adenovirus 2] >gi|57390...   195   8e-49
gi|109081029|ref|XP_001113023.1|  PREDICTED: dUTP pyrophosphatase   186   4e-46
gi|109081023|ref|XP_001112992.1|  PREDICTED: dUTP pyrophosphatase   186   4e-46
gi|109081025|ref|XP_001113050.1|  PREDICTED: dUTP pyrophosphatase   186   5e-46
gi|109081031|ref|XP_001112963.1|  PREDICTED: dUTP pyrophosphatase   182   5e-45
gi|78068021|gb|ABB18344.1|  ORF1 [Fowl adenovirus 10]               182   5e-45
gi|7513966|pir||T10819  dUTP diphosphatase (EC 3.6.1.23) - rat      181   1e-44
gi|94400930|ref|NP_001035361.1|  deoxyuridine triphosphatase isof   181   1e-44  G
gi|16758374|ref|NP_446044.1|  deoxyuridine triphosphatase isof...   181   1e-44  G
gi|21708114|gb|AAH33645.1|  DUTP pyrophosphatase [Homo sapiens...   180   2e-44  G
gi|60654529|gb|AAX29955.1|  dUTP pyrophosphatase [synthetic const   180   2e-44
gi|60831844|gb|AAX36986.1|  dUTP pyrophosphatase [synthetic const   180   2e-44
gi|78068007|gb|ABB18331.1|  ORF1 [Fowl adenovirus 4]                180   3e-44
gi|47123503|gb|AAH70339.1|  DUTP pyrophosphatase, isoform 1 pr...   179   3e-44  G
gi|70906444|ref|NP_001020420.1|  dUTP pyrophosphatase isoform ...   179   3e-44  G
gi|3041664|sp|P33316|DUT_HUMAN  Deoxyuridine 5'-triphosphate n...   179   3e-44  G
gi|34810582|pdb|1Q5U|Z  Chain Z, Human Dutp Pyrophosphatase >g...   179   3e-44  S
gi|74000610|ref|XP_851889.1|  PREDICTED: similar to Deoxyuridi...   179   6e-44  G
gi|55236363|gb|EAA13891.2|  ENSANGP00000012604 [Anopheles gamb...   178   7e-44  G
gi|84201701|gb|AAI11642.1|  Unknown (protein for MGC:137641) [Bos   178   7e-44
gi|41018627|gb|AAR98232.1|  ORF007 dUTPase [Orf virus] >gi|410...   178   1e-43  G
gi|18043437|gb|AAH19979.1|  Dutp protein [Mus musculus] >gi|74...   176   3e-43  G
gi|21281687|ref|NP_076084.2|  deoxyuridine triphosphatase [Mus...   176   3e-43  G
gi|53130842|emb|CAG31750.1|  hypothetical protein [Gallus gallus]   176   3e-43
gi|8347092|gb|AAF74514.1|  dUTPase [Mus musculus]                   176   3e-43  G
gi|50752967|ref|XP_413814.1|  PREDICTED: similar to bumetanide...   176   3e-43  G
gi|47230657|emb|CAF99850.1|  unnamed protein product [Tetraodon n   175   6e-43
```

Figure 2-27: NCBI BLAST output form: best match listing.



Query sequence

Matching sequence

Database sequence

```
> □gi|56160739|ref|AP_000401.1|  ORF1 [Fowl adenovirus 1]
 gi|9628836|ref|NP_043869.1| G dUTP pyrophosphatase [Fowl adenovirus A]
 gi|58424|emb|CAA78921.1|  unnamed protein product [Fowl adenovirus 1]
 gi|1314433|gb|AAC54895.1| G ORF1
 gi|7531092|sp|Q89662|DUT_ADEG1  Deoxyuridine 5'-triphosphate nucleotidohydrolase
pyrophosphatase)
Length=178

 Score =  361 bits (927),  Expect = 6e-99
 Identities = 178/178 (100%), Positives = 178/178 (100%), Gaps = 0/178 (0%)

Query  1    MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA  60
            MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA
Sbjct  1    MDPFGSSSVPPCSTSDLPEPKLYFVRLSPHAVPPVRATHGAAGYDLFSAYDIKVPARGRA  60

Query  61   LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR  120
            LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR
Sbjct  61   LVPTDLVFQFPPGCYGRIAPRSGLAAKFFIDVGAGVIDPDYRGNVSVVLFNFSESSFNIR  120

Query  121  RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR  178
            RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR
Sbjct  121  RGDRVAQLILERIMVPELSELTQLGETDRGASGFGSTGMGAVDRNQRSVLEWLTPGSR  178
```

Figure 2-28: NCBI BLAST output form: local sequence alignments.

Making a Multiple Protein Sequence Alignment with ClustalW

Besides running database searches to identify similar proteins pair by pair, the second most common bioinformatic task that biologists like to perform with protein sequences is a multiple alignment. *Multiple alignments* consist in lining up many similar proteins side by side for the sake of comparison. Multiple alignments are used to

- Identify sequence positions where specific amino acids really matter for the structural integrity or the function of a given protein
- Define specific sequence signatures for protein families
- Classify sequences and build evolutionary trees

We detail all the intricacies of making good multiple alignments in Chapter 9. At this point, however, we simply want to show you that performing a multiple alignment is really easy — especially when you have the pleasure of using some nice Internet server, such as the one maintained by the Protein Information Resource (PIR) people at Washington, D.C.'s Georgetown University.

The PIR actually originated from the *Atlas of Protein Sequences,* the first protein-sequence collection (which was built by the late Prof. M. Dayhoff in the late 1970s). The PIR site offers some useful protein analysis tools and databases that we invite you to explore by yourself. Among these tools, it offers a multiple-alignment server (running the standard ClustalW program) that is really easy to use for beginners.

Do the following to get your feet wet using ClustalW:

1. **Point your browser to** `pir.georgetown.edu`**.**

   The PIR home page appears, as shown in Figure 2-29.

2. **Under the Search/Analysis heading, choose Multiple Alignment from the drop-down menu to display the input form.**

   The input form appears. At this point, you need a few FASTA-formatted protein sequences.

3. **Open the dUTPase FASTA-formatted sequence file that you created on your PC in the previous section of this chapter.**

Alternatively, you can go to the Swiss-Prot home page at

`www.expasy.org/sprot/`

and grab a few dUTPase sequences (such as `DUT_AQUAE`, `DUT_BRAJA`, `DUT_BPT5`, `DUT_CANAL`, and `DUT_BUCAI`), as we spell out in the section “ExPASy: A prime Internet site for protein information,” earlier in this chapter.

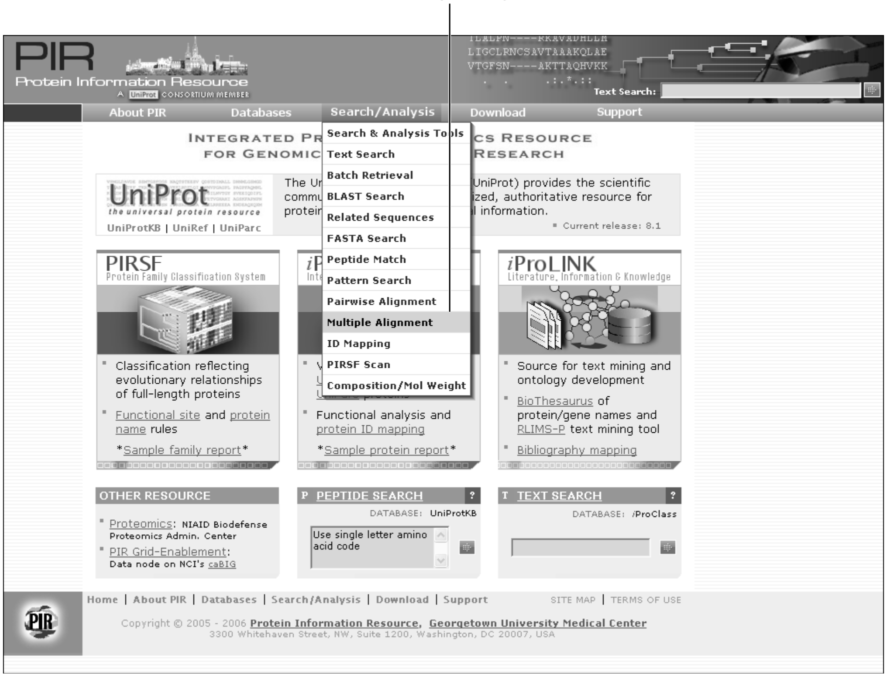
4. **Copy/paste five of these sequences (all at once) into the input window, as shown in Figure 2-30.**

   Be careful to copy the definition line and the *entire* amino-acid sequence for each of them.

5. **Click the Submit button.**

   Within a few seconds, your screen (Figure 2-31) is proudly displaying its first successful multiple-sequence alignment — and a tree-like representation of the pair-wise similarity of these sequences.

Select Multiple Alignment.



The Un… …UniProt) provides the scientific commu… …ized, authoritative resource for protein… …l information.

Current release: 8.1

Classification reflecting evolutionary relationships of full-length proteins

Functional site and protein name rules

Sample family report

Functional analysis and protein ID mapping

Sample protein report

Source for text mining and ontology development

BioThesaurus of protein/gene names and RLIMS-P text mining tool

Bibliography mapping

Proteomics: NIAID Biodefense Proteomics Admin. Center

PIR Grid-Enablement: Data node on NCI's caBIG

Copyright © 2005 - 2006 Protein Information Resource, Georgetown University Medical Center
3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, USA

Figure 2-29: Another useful protein analysis center: PIR.

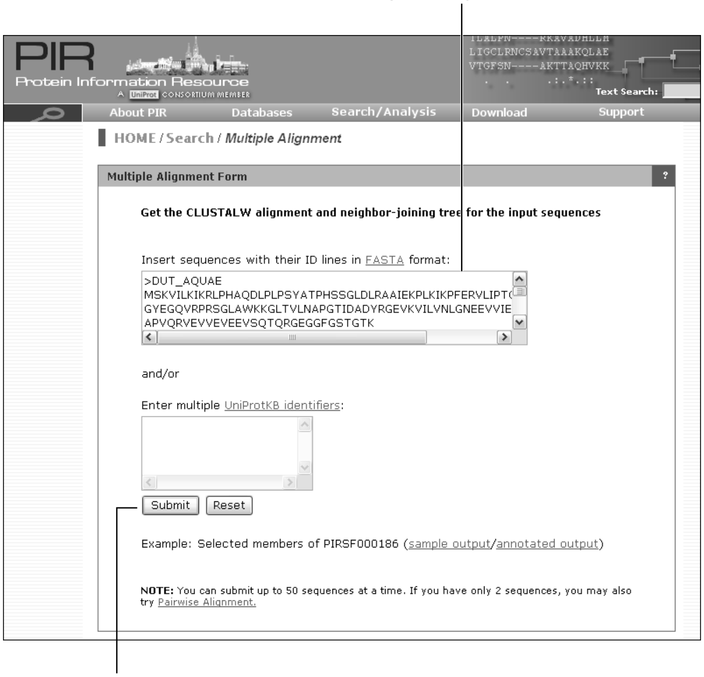
This output (if you used the same sequences as we did here) illustrates the power of multiple alignments:

- It clearly delineates a region of high-versus-low sequence similarity in these related proteins.

  Positions 100 percent identical (highlighted by <*>) or occupied by chemically similar amino-acids (highlighted by <:> or <.>) tend to occur in a clustered fashion along the sequence, defining regions that are probably directly involved in shaping the catalytic site or performing the pyrophosphatase biochemical reaction.
- Various types of *functional signatures* (small sequence segments associated to a given biochemical function) can be extracted from such a multiple alignment.
- Evolutionary relationships between proteins are inferred from phylogenetic trees.

In Chapters 9, 11, and 13, we tell you more about these key applications of bioinformatics.

Paste your sequences here.



Get the CLUSTALW alignment and neighbor-joining tree for the input sequences

Insert sequences with their ID lines in FASTA format:

```
>DUT_AQUAE
MSKVILKIKRLPHAQDLPLPSYATPHSSGLDLRAAIEKPLKIKPFERVLIPT(
GYEGQVRPRSGLAWKKGLTVLNAPGTIDADYRGEVKVILVNLGNEEVVIE
APVQRVEVVEVEEVSQTQRGEGGFGSTGTK
```

and/or

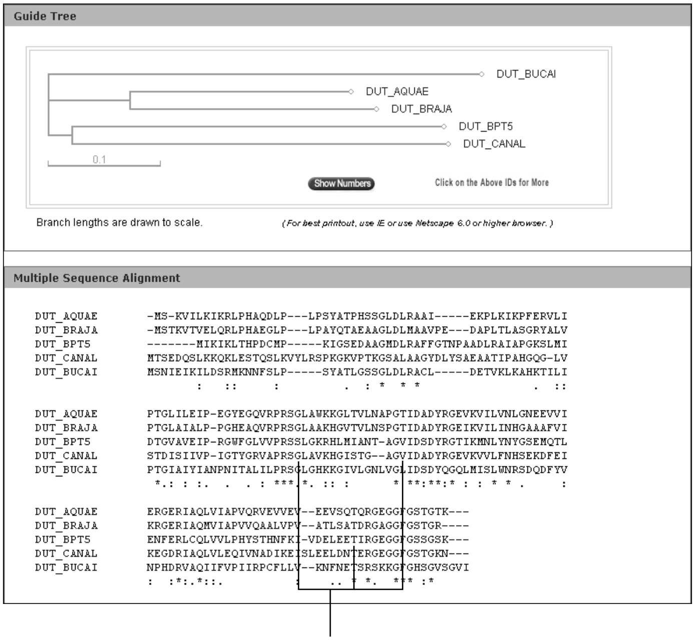
Enter multiple UniProtKB identifiers:

Example: Selected members of PIRSF000186 (sample output/annotated output)

NOTE: You can submit up to 50 sequences at a time. If you have only 2 sequences, you may also try Pairwise Alignment.

Click here to start the multiple alignment.

Figure 2-30: Filling up the multiple alignment input window at PIR.



Branch lengths are drawn to scale.

(For best printout, use IE or use Netscape 6.0 or higher browser.)

Multiple Sequence Alignment

```
DUT_AQUAE      -MS-KVILKIKRLPHAQDLP---LPSYATPHSSGLDLRAAI-----EKPLKIKPFERVLI
DUT_BRAJA      -MSTKVTVELQRLPHAEGLP---LPAYQTAEAAGLDLMAAVPE---DAPLTLASGRYALV
DUT_BPT5       -------MIKIKLTHPDCMP-----KIGSEDAAGMDLRAFFGTNPAADLRAIAPGKSLMI
DUT_CANAL      MTSEDQSLKKQKLESTQSLKVYLRSPKGKVPTKGSALAAGYDLYSAEAATIPAHGQG-LV
DUT_BUCAI      MSNIEIKILDSRMKNNFSLP-----SYATLGSSGLDLRACL-----DETVKLKAHKTILI
                      :   ::     :         .  : *  * *                .  ::

DUT_AQUAE      PTGLILEIP-EGYEGQVRPRSGLAWKKGLTVLNAPGTIDADYRGEVKVILVNLGNEEVVI
DUT_BRAJA      PTGLAIALP-PGHEAQVRPRSGLAAKHGVTVLNSPGTIDADYRGEIKVILINHGAAAFVI
DUT_BPT5       DTGVAVEIP-RGWFGLVVPRSSLGKRHLMIANT-AGVIDSDYRGTIKMNLYNYGSEMQTL
DUT_CANAL      STDISIIVP-IGTYGRVAPRSGLAVKHGISTG--AGVIDADYRGEVKVVLFNHSEKDFEI
DUT_BUCAI      PTGIAIYIANPNITALILPRSGLGHKKGIVLGNLVGLIDSDYQGQLMISLWNRSDQDFYV
                *.: : :.  .  . : ***.*. :: :      * **:**:* : : * * .     :

DUT_AQUAE      ERGERIAQLVIAPVQRVEVVEV--EEVSQTQRGEGGFGSTGTK---
DUT_BRAJA      KRGERIAQMVIAPVVQAALVPV--ATLSATDRGAGGFGSTGR----
DUT_BPT5       ENFERLCQLVVLPHYSTHNFKI-VDELEETIRGEGGFGSSGSK---
DUT_CANAL      KEGDRIAQLVLEQIVNADIKEISLEELDNTERGEGGFGSTGKN---
DUT_BUCAI      NPHDRVAQIIFVPIIRPCFLLV--KNFNETSRSKKGFGHSGVSGVI
               :  :*:.*::.          :    .. * *.  *** :*
```

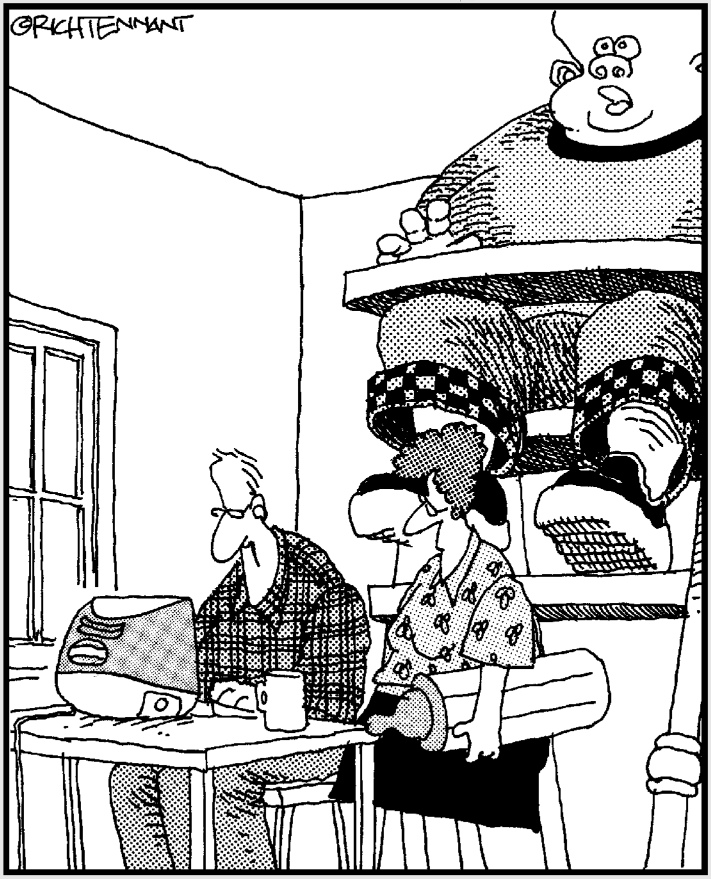
Invariant regions

Figure 2-31: Multiple alignment of five bacterial dUTPase sequences and the corresponding tree built from their pair-wise similarity.

Part II

A Survival Guide to Bioinformatics

The 5th Wave By Rich Tennant



"Well, that proves it. He's definitely from your side of the family."

In this part . . .

Bioinformatics is about finding and interpreting biological data online. In this part, we show you how to go shopping in the main bioinformatics databases. We've got lots of great tips for picking up the freshest, ripest data (and not just the low-hanging fruits!). Because all these databases are linked with one another, we also show you how you can travel across them and gather every piece of information you need. It's a fascinating journey across human knowledge — and a compulsory starting point for any research project.

With the right data in the fridge, it's time to start cooking. In this part, we show simple recipes for working with DNA and protein sequences in order to predict their most basic properties.

Chapter 3

Using Nucleotide Sequence Databases

In This Chapter

- Nabbing a quick refresher on the structure of genes and genomes
- Making use and sense of GenBank
- Finding out about a specific gene
- Working with complete microbial genomes
- Browsing the human and other animal genomes

> *The secret of success is to know something nobody else knows.*
>
> — Aristotle Onassis (1906–1975)

Sequence databases are great tools because they offer a unique window on the past. They make it possible to answer today's biological questions by enabling us to analyze sequences that may have been determined as many as 25 years ago, when the whole technology emerged. By doing this, they connect past and present molecular biology.

The first databases were in fact created as some sort of sequence museum, where sequences could be preserved for all eternity in pristine form, just as they were determined, interpreted, and published by their original authors. This *historical* (time capsule!) perspective pretty much remains in GenBank, the leading nucleotide sequence repository maintained as a consortium between the U.S. National Center for Biotechnology Information (NCBI), the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ). In this chapter, we show you how to use GenBank and decipher its entries.

Repository-type databases are great tools when you want to come up with a bibliography for a particular sequence, but they do not provide easy access to sequence data when your query deals with broader issues related to a gene or function rather than with a specific paper. For this reason, the second-generation nucleotide-sequence databases have adopted a more *gene-centric* perspective, where all the sequence information relevant to a given gene is made accessible at once. In addition to unraveling the mysteries of the more traditional GenBank, then, we also show you how to use NCBI's Entrez Gene, a good example of a gene-centric database.

These days, nucleotide sequences are routinely determined at the whole-genome or chromosome scale — at least for microorganisms. We now have information not only about individual gene sequences, but also about their relative positions, strand orientation, and the presence or absence of biochemical functions within an entire organism. To take advantage of this more global information, researchers have had to design state-of-the-art *genome-centric* sequence-information management systems that can connect specialized sequence collections with browsing tools. As an added bonus, this chapter presents some of the great genome-centric resources dealing with viruses, bacteria, or (you guessed it) human beings.

However, before you start delving into these information-management systems in earnest, you may find it useful to read the next section. There, we quickly summarize the fundamentals of genes and genomes and introduce the vocabulary you need to read GenBank fluently. If you're an experienced biologist, you can skip this section completely — or read it word for word in hopes of finding some embarrassing mistake. Go ahead. We dare you! Conversely, if you want to learn much more, we advise you to go get *Genetics For Dummies,* by Tara Rodden Robinson, a great companion book for would-be bioinformaticians.

Reading into Genes and Genomes

True, nucleotide sequences are universal, but the structure of the genes they encode is markedly different between *prokaryotic* organisms (organisms lacking a true nucleus) and *eukaryotic* organisms (the kind lucky enough to have a true nucleus). You have to know the basic architecture of both types of genomes and genes to make sense of a simple GenBank entry. Remember that we need to define a *gene* as the contiguous genome segment encompassing all the nucleotide-sequence information necessary to bring about its successful *expression* — that is, the production of protein or RNA. This handy definition includes both the coding and regulatory parts of a gene.

Prokaryotes: Small bugs, simple genes

The three most basic classes of living organisms are the *prokaryotes* (usually bacteria), the *archaea* (bacteria-like organisms living in extreme conditions), and the *eukaryotes.* Eukaryotes go all the way from microscopic yeast to humans, animals, and plants.

For bioinformatic purposes, prokaryotes and archaea are very much the same. With few exceptions — far beyond the realm of this book — they have the following properties in common:

- They are microscopic organisms.
- Their genome is a single, circular DNA molecule.
- Their genome size is in the order of a few million base pairs [0.6–8].
- Their *gene density* — the number of genes per base pairs in the genome — is approximately one gene per 1,000 base pairs.
- Their genome is lean and mean, containing few useless parts (70 percent is coding for proteins).
- Their genes do not overlap.
- Their genes are transcribed (copied into messenger RNA) right after a control region called a *promoter.*
- These messenger RNA (mRNA) are collinear with the genome sequence. In other words, genes are in a single piece, not interrupted by noncoding patches (called *introns*).
- Protein sequences are derived by translating the longest open reading frame (from ATG to STOP) spanning the gene-transcript sequence.

Figure 3-1 illustrates the simple collinear relationship between a bacterial genomic (gene) sequence, the transcript (mRNA), the open reading frame (ORF), and the final protein. (For a peek at a typical [small] bacterial genome, flip to Figure 1-10 in Chapter 1.) The mRNA sequence gets translated into a protein right after a special signal, called the Ribosome Binding Site (RBS in Figure 3-1). The ribosome is the main piece of machinery of the cell's protein-translation apparatus.

Accordingly, database entries describing a coding prokaryotic sequence should include three important features:

- The coordinates of some promoter elements
- The coordinates of the RBS
- The coordinates of the ORF boundaries

That's about all there is to say regarding the simple architecture of prokaryotic genes. Not all genes encode proteins; for some of them, the function is directly carried out by the transcribed RNA molecule. This includes transfer RNA (tRNA), ribosomal RNA (rRNA), and a few other fancy RNAs that shall remain nameless.

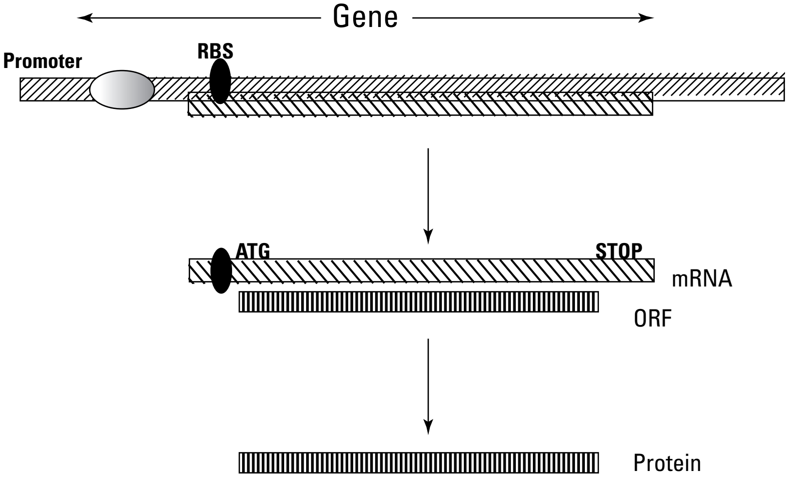


Figure 3-1: Relationship between gene, mRNA, and protein sequence for prokaryotes.

Eukaryotes: Bigger bugs, complex genes

Eukaryotes encompass a wide variety of organisms, from microscopic yeasts and fungi to elephants, plants, and humans. Yet eukaryotes, too, have basic properties in common — properties that end up making them a bit more challenging when it comes to genomic and bioinformatic analysis:

- Their genome consists of multiple linear pieces of DNA called *chromosomes* (up to a hundred million base pairs long).
- Their genome size (10–670,000 million base pairs), especially for animals, plants, and amoebae (!), is *much bigger* than in prokaryotes.
- Their gene density is much lower than that for prokaryotes (one human gene per 100,000 base pairs).
- Their genome is no model of efficiency, containing many useless parts (less than 5 percent of the human genome code for proteins).
- Genes on opposite DNA strands might overlap — although that's a relatively rare occurrence.
- Their genes are transcribed right after a control region called a *promoter,* but sequence elements located far away can have a strong influence on this process.
- Gene sequences are not collinear with the final messenger RNA (mRNA) and protein sequences. Only small bits (the *exons*) are retained in the mature mRNA that encodes the final product.
- Genes often exhibit more than one mRNA (and protein) form.

Genes of higher eukaryotes (animals) may span up to millions of base pairs — the human dystrophin gene (the mutation of which causes a dreadful disease), for example, is 2.2 million base pairs long. The relationships among a gene DNA sequence, its primary transcript, the various forms of mature mRNA, and the final protein sequence can be very complex. This is why understanding a database entry that describes all these things may take a bit of work. In addition, a lot of database entries correspond to partial gene sequences, or different gene-related objects (such as promoter regions, mRNA, or genome fragments). Throughout the rest of this chapter, we show you several examples so that, after a bit of study, you can decipher the most intricate GenBank entries by yourself.

Making Use (and Sense) of GenBank

For prokaryotes, the limited size of the genes involved — as well as the simple (linear) relationship among the gene DNA sequence, the mRNA, the ORFs, and the final protein sequence — make all that information relatively easy to annotate and store in database records. This is why database entries corresponding to bacterial genes are relatively easy to read and understand. In this section, we take you through the entry of the *Escherichia coli* dUTPase gene step by step (handy GenBank ID X01714).

Making sense of the GenBank entry of a prokaryotic gene

First things first, we have to fetch our sample GenBank entry, as follows:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   The NCBI PubMed home page appears.

2. **From the Search pull-down menu, choose Nucleotide (as in Figure 3-2).**

3. **Type the GenBank ID X01714 in the For field, and then click Go.**

   A Results page appears, displaying a short definition for the nucleotide sequence entry, preceded by its ID X01714 (underlined in blue).

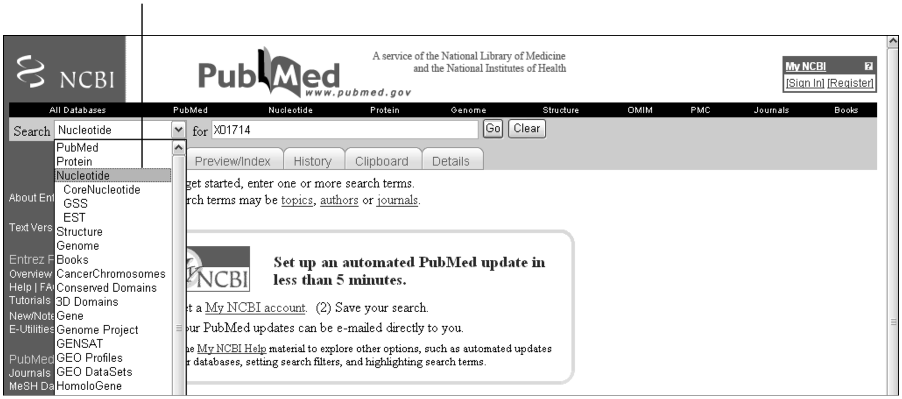
4. **Click the X01714 hyperlink. (Alternatively, you can change the display option from Summary to GenBank.)**

   The X01714 GenBank entry appears in the default GenBank format, as shown in Figure 3-3.

   Gurus refer to this as the GenBank *flat-file format* because you can read it in a linear fashion, and it doesn't involve indexes, pointers, or accessory information (which are customary in the structure of more sophisticated databases).

   In fact, what you have here isn't totally "flat" because it contains a few hyperlinks.

5. **Select the Text option from the Send To pull-down menu to generate a true flat file (text) format of the X01714 entry.**
6. **Save your entry as a text document on your hard drive by choosing File➪Save As from your browser's main menu.**



get started, enter one or more search terms.
rch terms may be topics, authors or journals.

Set up an automated PubMed update in less than 5 minutes.

t a My NCBI account. (2) Save your search.
ur PubMed updates can be e-mailed directly to you.
e My NCBI Help material to explore other options, such as automated updates
r databases, setting search filters, and highlighting search terms.

Choose Nucleotide.

Figure 3-2: Searching for nucleotide sequence X01714 at NCBI PubMed.

Reading the header of a GenBank prokaryotic entry

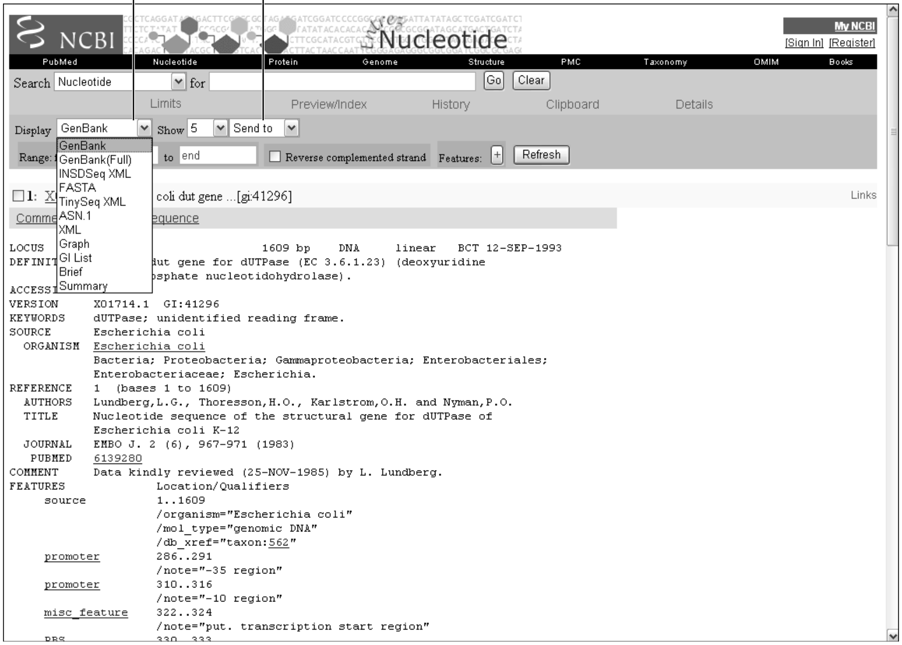
The entry text saved in the preceding steps is divided into sections, with keywords introducing each section. Typical keywords include LOCUS, DEFINITION, ACCESSION, VERSION, KEYWORDS, and SOURCE. Refer to Figure 3-3 as we go through these keywords step by step:

- **LOCUS** gives us the locus name (an arbitrary name; here it's ECDUT), the size of the nucleotide sequence in base pairs, the nature of the molecule (here it's DNA), and its topology (*linear* or *circular* molecule — in this particular case, linear).
- **DEFINITION** provides a short definition of the gene that corresponds to the entry sequence. Here, it's the *E. coli* dUTPase gene.
- **ACCESSION** lists the accession number — a unique identifier within and across various databases (such as the protein-sequence databases). Here, the accession number is X01714.
- **VERSION** fills you in on synonymous or past ID numbers.
- **KEYWORDS** introduces a list of terms that broadly characterize the entry. You can use these terms as keywords for certain database searches. (For more on using keywords in database searches, check out Chapter 2, where we discuss using PubMed with limits.)
- **SOURCE** divulges the common name of the relevant organism to which the sequence belongs.

- **ORGANISM** gives a more complete identification of the organism, complete with its technical taxonomic classification.

  Notice that ORGANISM is set in a bit from the left margin of the page — more so, at least, than the other keywords. This isn't a mistake; computer scientists call this arrangement *indentation.* The precise position where a keyword begins on the various lines denotes its subordination to other keywords. Keywords starting at the same position are said to be at the same level. The leftmost position delimits sections of the GenBank entry that are independent, such as the preceding six sections.

  In contrast, ORGANISM is indented to indicate that it's part of the SOURCE section. (The same thing happens in the REFERENCE section.)
- **REFERENCE** introduces a section where the credits for the sequence determination are given (different parts of the sequences can be credited to different authors). The REFERENCE section contains multiple parts as denoted by the indentation of the following self-explanatory keywords: AUTHORS, TITLE, JOURNAL, and PUBMED.
- **COMMENT** introduces the last line of the top of the X01714 GenBank entry. It contains free-formatted text, such as acknowledgments or info that doesn't fit in the previous sections.



Use this menu for FASTA format.

Click here for text format.

```
LOCUS                                  1609 bp    DNA     linear   BCT 12-SEP-1993
DEFINIT                dut gene for dUTPase (EC 3.6.1.23) (deoxyuridine
                       osphate nucleotidohydrolase).
ACCESSI
VERSION     X01714.1  GI:41296
KEYWORDS    dUTPase; unidentified reading frame.
SOURCE      Escherichia coli
  ORGANISM  Escherichia coli
            Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
            Enterobacteriaceae; Escherichia.
REFERENCE   1  (bases 1 to 1609)
  AUTHORS   Lundberg,L.G., Thoresson,H.O., Karlstrom,O.H. and Nyman,P.O.
  TITLE     Nucleotide sequence of the structural gene for dUTPase of
            Escherichia coli K-12
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /mol_type="genomic DNA"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc_feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
```

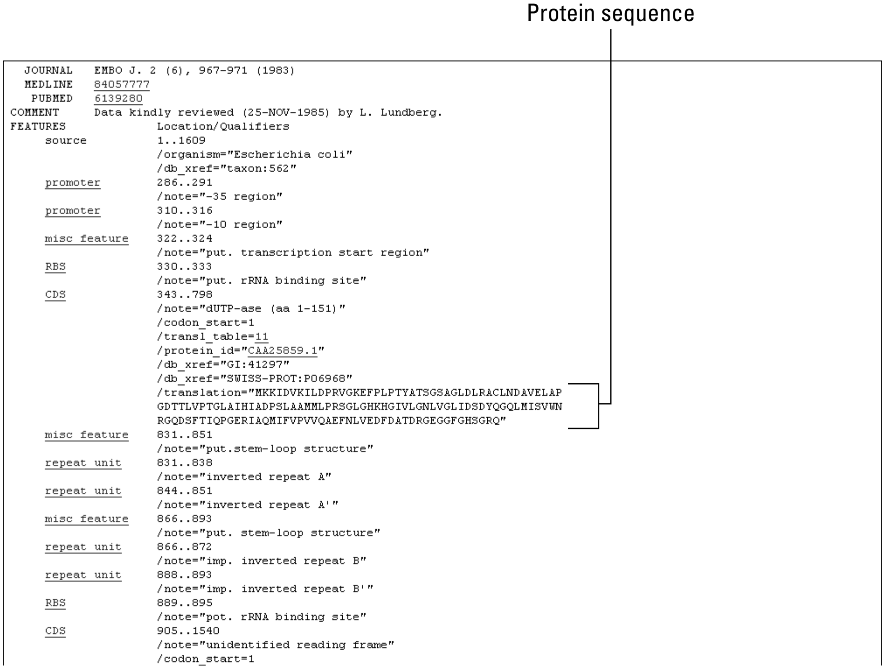
Figure 3-3: GenBank entry X01714.

The next (middle) section of the entry is the *FEATURES table* (see Figure 3-4). It describes precisely the gene regions and the associated biological properties that have been identified in the nucleotide sequence. This entire section is under the control of the FEATURES keyword. A large variety of biologically related keywords are subordinated to FEATURES, as we make clear in the following list:

- **source** indicates the origin of specific regions of the sequence. This is useful when you want to distinguish cloning vectors from host sequences. In X01714, the whole sequence comes from *E. coli* genomic DNA.
- **promoter** shows the precise coordinates of a promoter element. In X01714, a *–35 region* is indicated from position 286 to 291 (indicated by *286..291*) in the nucleotide sequence.

  promoter introduces another line that contains the coordinates of another promoter element, this time a *–10 region* at positions 310 to 316.
- **misc feature** (miscellaneous feature) indicates the putative location of the transcription start (mRNA synthesis). For X01714, this is from positions 322 to 324.
- **RBS** (Ribosome Binding Site) indicates the location of the last upstream element. For X01714, this is at position 330 to 333.
- **CDS** (CoDing Segment) introduces a complex section that describes the gene's *open reading frame (ORF)*:
  - The first line indicates the coordinates of the ORF from its initial **A**TG to the last nucleotide of the first stop codon TA**A**. (For X01714, it's 343 to 798.) See Chapter 1 if you don't remember what a codon is.
  - Each of the following lines (indented at the same level) gives the name of a protein product, indicates the reading frame to use (here, 343 is the first base of the first codon), the genetic code to apply, and a number of IDs for the protein sequence.
  - `/translation`, the final keyword of the CDS section, introduces the conceptual amino-acid sequence of the coding segment. This sequence is a computer translation that uses the coordinates, reading frame, and genetic code indicated in the preceding lines.
- Another **misc feature** follows the CDS section. It contains lines that point out putative stem-loop structures and repeats. These are potential regulatory elements of the dUTPase gene.

The FEATURES section of entry X01714 is typical of a simple, well-annotated bacterial nucleotide sequence, centered on a well-identified gene. However, this straightforward entry includes a complication: a putative additional reading frame. We selected this entry because we were interested in one gene: the dUTPase. However, the entry exhibits an extra putative gene, indicated by an additional RBS element and a second CDS section.



Protein sequence

```
  JOURNAL   EMBO J. 2 (6), 967-971 (1983)
  MEDLINE   84057777
   PUBMED   6139280
COMMENT     Data kindly reviewed (25-NOV-1985) by L. Lundberg.
FEATURES             Location/Qualifiers
     source          1..1609
                     /organism="Escherichia coli"
                     /db_xref="taxon:562"
     promoter        286..291
                     /note="-35 region"
     promoter        310..316
                     /note="-10 region"
     misc feature    322..324
                     /note="put. transcription start region"
     RBS             330..333
                     /note="put. rRNA binding site"
     CDS             343..798
                     /note="dUTP-ase (aa 1-151)"
                     /codon_start=1
                     /transl_table=11
                     /protein_id="CAA25859.1"
                     /db_xref="GI:41297"
                     /db_xref="SWISS-PROT:P06968"
                     /translation="MKKIDVKILDPRVGKEFPLPTYATSGSAGLDLRACLNDAVELAP
                     GDTTLVPTGLAIHIADPSLAAMMLPRSGLGHKHGIVLGNLVGLIDSDYQGQLMISVWN
                     RGQDSFTIQPGERIAQMIFVPVVQAEFNLVEDFDATDRGEGGFGHSGRQ"
     misc feature    831..851
                     /note="put.stem-loop structure"
     repeat unit     831..838
                     /note="inverted repeat A"
     repeat unit     844..851
                     /note="inverted repeat A'"
     misc feature    866..893
                     /note="put. stem-loop structure"
     repeat unit     866..872
                     /note="imp. inverted repeat B"
     repeat unit     888..893
                     /note="imp. inverted repeat B'"
     RBS             889..895
                     /note="pot. rRNA binding site"
     CDS             905..1540
                     /note="unidentified reading frame"
                     /codon_start=1
```

Figure 3-4: GenBank entry X01714: the FEATURES table part.

GenBank entries containing more than one gene are frequent.

Using the Sequence section of a prokaryotic entry

The last section of GenBank entry X01714 is the nucleotide sequence section. It starts with the *ORIGIN* keyword and finishes with the end-of-entry line introduced by two slash marks (//). Each line of nucleotide sequence starts with the position number of the first nucleotide in that line. Each line contains 60 nucleotides.

Because the sequence section of a GenBank entry mixes numbers and nucleotides, you *cannot* directly use it as an input on most sequence analysis servers. You first need to generate a FASTA-formatted sequence, as follows:

1. **On your GenBank entry page, choose FASTA from the Display pull-down menu. (Refer to Figure 3-3.)**

   You'll find the Display pull-down menu at the top-left of the page.

2. **Select the Text option from the Send To pull-down menu.**

   The nucleotide sequence appears now in a form suitable for most sequence-analysis programs.

3. **Save this entry by choosing File⇨Save As from your browser's main menu or by copying/pasting it into an empty Word document.**

Making sense of the GenBank entry of an eukaryotic mRNA

We now continue with the dUTPase gene because it's quite ubiquitous and present in both prokaryotes and eukaryotes: Humans have it, too. Looking at the GenBank entries describing the human version of dUTPase clearly illustrates the increased complexity of higher eukaryotes compared to prokaryote genes. We start here with a simple one, GenBank entry U90223:

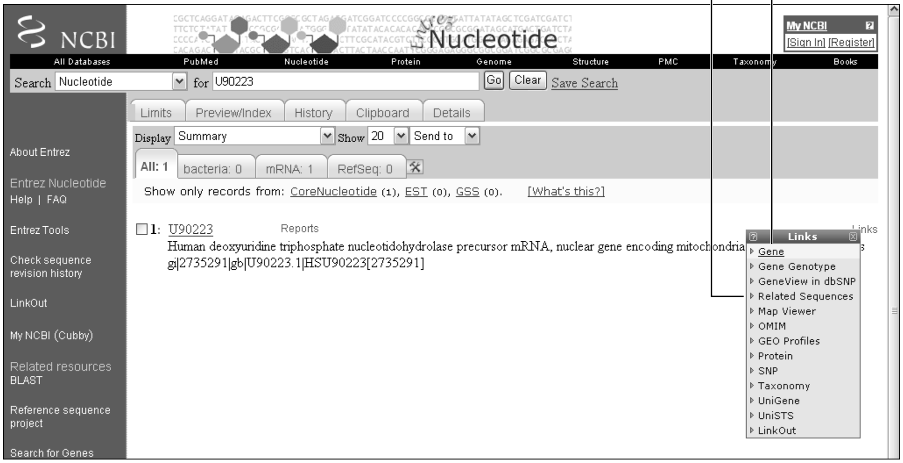
1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   The NCBI PubMed home page appears on cue.

2. **From the Search pull-down menu, choose Nucleotide.**

3. **Type in GenBank ID U90223 in the For window, and click Go.**

   A Results page appears, displaying a short definition for the nucleotide sequence entry, preceded by its ID U90223 (underlined in blue) as shown in Figure 3-5.



Jump to the corresponding gene.

Get related sequences.

Search Nucleotide for U90223

Display Summary Show 20

□1: U90223 Reports
Human deoxyuridine triphosphate nucleotidohydrolase precursor mRNA, nuclear gene encoding mitochondria
gi|2735291|gb|U90223.1|HSU90223[2735291]

Figure 3-5: Initial display for GenBank entry U90223.

4. **Click the U90223 hyperlink. (Alternatively, you can change the Display option from Summary to GenBank.)**

   You're now ready to read your first human GenBank entry.

5. **Select the Text option from the Send To pull-down menu to generate a true flat-file (text-format) entry.**
6. **Save your result by choosing File➪Save As.**

Although it's related to a human gene, GenBank entry U90223 doesn't look very different from entry X01714, the one that describes its bacterial homologue. The top part of the entry follows the general information keywords order: LOCUS, ACCESSION, DEFINITION, and VERSION.

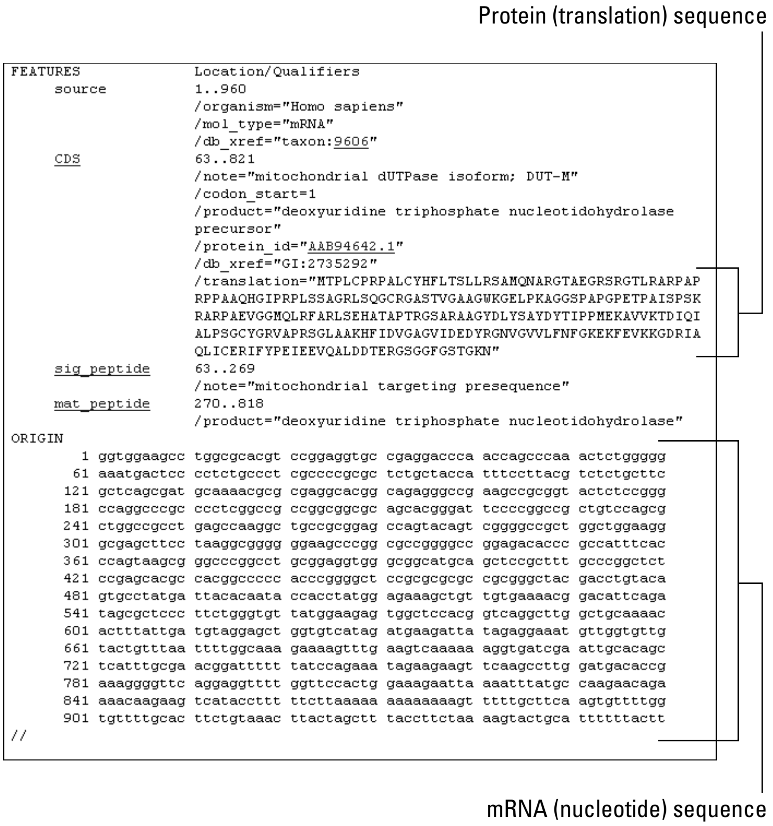
The KEYWORD line, which is supposed to list readily relevant and searchable terms (such as dUTPase), is empty for entry U90223. Unfortunately, this isn't a fluke. It illustrates a common problem in sequence databases — annotations may be incomplete. As a result, keyword-based database searches usually don't retrieve all the relevant sequences in GenBank. On a related note, the information slated for the SOURCE and REFERENCE sections may also be missing or incomplete. The U90223 entry, for example, says that the sequence here comes from an unpublished source. However, a simple PubMed search (see Chapter 2) using the author's names — Ladner RD and Caradonna SJ — brings us the relevant article in the *Journal of Biological Chemistry* — from 1997! A word to the wise: You should never expect GenBank (or any sequence database) annotations to be up to date; no wonder it's sometimes impossible to retrieve some sequences with PubMed searches.

Figure 3-6 displays the FEATURES section of the U90223 entry. The CDS keyword indicates a coding region (63-821) sequence that corresponds to the mitochondrial form of human dUTPase. Following the conceptual amino-acid translation of the ORF, the *sig peptide* keyword indicates the location of a mitochondrial targeting sequence, and the *mat peptide* keyword provides the exact boundaries of the mature peptide.

This human GenBank entry isn't really more complex than its bacterial homologue because we carefully selected an entry describing an *mRNA* sequence, not a *genomic* one. At the level of the mature mRNA, the relationship between a eukaryotic protein and its encoding nucleotide sequences is as collinear as its prokaryotic counterpart. As the next section shows, genuine genomic data can be a bit trickier.

Making sense of a GenBank eukaryotic genomic entry

The previous section described the GenBank entry corresponding to the mRNA sequence of the human dUTPase gene. In this case, we want to look at the GenBank entry AF018430 related to the *gene sequence* (as it is on the chromosome) from which this mRNA originated.



```
FEATURES             Location/Qualifiers
     source          1..960
                     /organism="Homo sapiens"
                     /mol_type="mRNA"
                     /db_xref="taxon:9606"
     CDS             63..821
                     /note="mitochondrial dUTPase isoform; DUT-M"
                     /codon_start=1
                     /product="deoxyuridine triphosphate nucleotidohydrolase
                     precursor"
                     /protein_id="AAB94642.1"
                     /db_xref="GI:2735292"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     sig_peptide     63..269
                     /note="mitochondrial targeting presequence"
     mat_peptide     270..818
                     /product="deoxyuridine triphosphate nucleotidohydrolase"
ORIGIN
        1 ggtggaagcc tggcgcacgt ccggaggtgc cgaggaccca accagcccaa actctggggg
       61 aaatgactcc cctctgccct cgccccgcgc tctgctacca tttccttacg tctctgcttc
      121 gctcagcgat gcaaaacgcg cgaggcacgg cagagggccg aagccgcggt actctccggg
      181 ccaggcccgc ccctcggccg ccggcggcgc agcacgggat tccccggccg ctgtccagcg
      241 ctggccgcct gagccaaggc tgccgcggag ccagtacagt cggggccgct ggctggaagg
      301 gcgagcttcc taaggcgggg ggaagcccgg cgccggggcc ggagacaccc gccatttcac
      361 ccagtaagcg ggcccggcct gcggaggtgg gcggcatgca gctccgcttt gcccggctct
      421 ccgagcacgc cacggccccc acccggggct ccgcgcgcgc cgcgggctac gacctgtaca
      481 gtgcctatga ttacacaata ccacctatgg agaaagctgt tgtgaaaacg gacattcaga
      541 tagcgctccc ttctgggtgt tatggaagag tggctccacg gtcaggcttg gctgcaaaac
      601 actttattga tgtaggagct ggtgtcatag atgaagatta tagaggaaat gttggtgttg
      661 tactgtttaa ttttggcaaa gaaaagtttg aagtcaaaaa aggtgatcga attgcacagc
      721 tcatttgcga acggattttt tatccagaaa tagaagaagt tcaagccttg gatgacaccg
      781 aaaggggttc aggaggtttt ggttccactg gaaagaatta aaatttatgc caagaacaga
      841 aaacaagaag tcataccttt ttcttaaaaa aaaaaaaagt ttttgcttca agtgttttgg
      901 tgttttgcac ttctgtaaac ttactagctt taccttctaa aagtactgca tttttactt
//
```

Figure 3-6: FEATURES section of GenBank entry U90223.

First, however, we have to fetch the entry by using the same protocol as before:

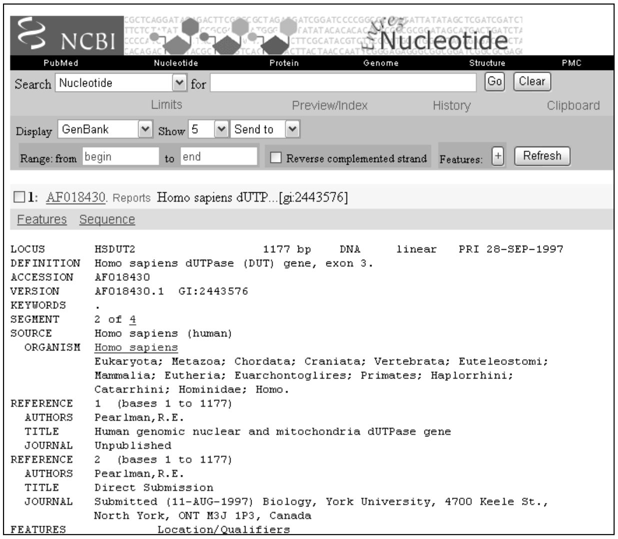
1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   The NCBI PubMed home page appears.

2. **From the Search pull-down menu, choose Nucleotide.**
3. **Type in AF018430 in the For window, and click Go.**
4. **Click the AF018430 link.**

   The AF018430 entry appears, as shown in Figure 3-7.

You're now ready to read your first human genomic GenBank entry. We've selected a simple one for you to start with, but it contains some of the specificities only encountered in eukaryotic entries.



Search Nucleotide for

Display GenBank Show 5 Send to

Range: from begin to end □ Reverse complemented strand Features: +

□ 1: AF018430. Reports Homo sapiens dUTP...[gi:2443576]

```
LOCUS       HSDUT2                  1177 bp    DNA     linear   PRI 28-SEP-1997
DEFINITION  Homo sapiens dUTPase (DUT) gene, exon 3.
ACCESSION   AF018430
VERSION     AF018430.1  GI:2443576
KEYWORDS    .
SEGMENT     2 of 4
SOURCE      Homo sapiens (human)
  ORGANISM  Homo sapiens
            Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
            Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
            Catarrhini; Hominidae; Homo.
REFERENCE   1  (bases 1 to 1177)
  AUTHORS   Pearlman,R.E.
  TITLE     Human genomic nuclear and mitochondria dUTPase gene
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 1177)
  AUTHORS   Pearlman,R.E.
  TITLE     Direct Submission
  JOURNAL   Submitted (11-AUG-1997) Biology, York University, 4700 Keele St.,
            North York, ONT M3J 1P3, Canada
FEATURES             Location/Qualifiers
```

Figure 3-7: GenBank entry AF018430.

The top part contains general information introduced by the usual keywords:

- **LOCUS:** The locus name is HSDUT2. The rest of the line tells us that we're dealing with 1177 bp of linear DNA.
- **DEFINITION:** This line indicates that the name of the gene is *DUT* and specifies that the entry encompasses *exon 3* of the gene. This reminds us that human genes generally spread their protein-coding regions across many disjointed pieces, the exons.
- The **ACCESSION**, **VERSION**, and **KEYWORDS** lines are standard. Note that for AF018430, the latter is empty, meaning you can't retrieve this entry via a keyword-field-restricted search.
- **SEGMENT:** This field relates to the mosaic structure of eukaryotic genes. It indicates that this current GenBank entry is the second segment of a super entry made of four. You need all four entries to reconstruct the complete mRNA sequence used as a template for producing the protein.
- The **ORGANISM**, **SOURCE**, and **REFERENCE** lines are standard.

The FEATURES table is what really makes a eukaryotic genomic entry special — and, as such, is much longer than the ones for prokaryotic organisms. It contains the following elements (see Figure 3-8):

- The **source** section contains a */map* section. For AF018430, it indicates that the sequence belongs to chromosome 15, and was more precisely mapped on the long arm (q) of this chromosome, within the q21.1 cytogenetic band.

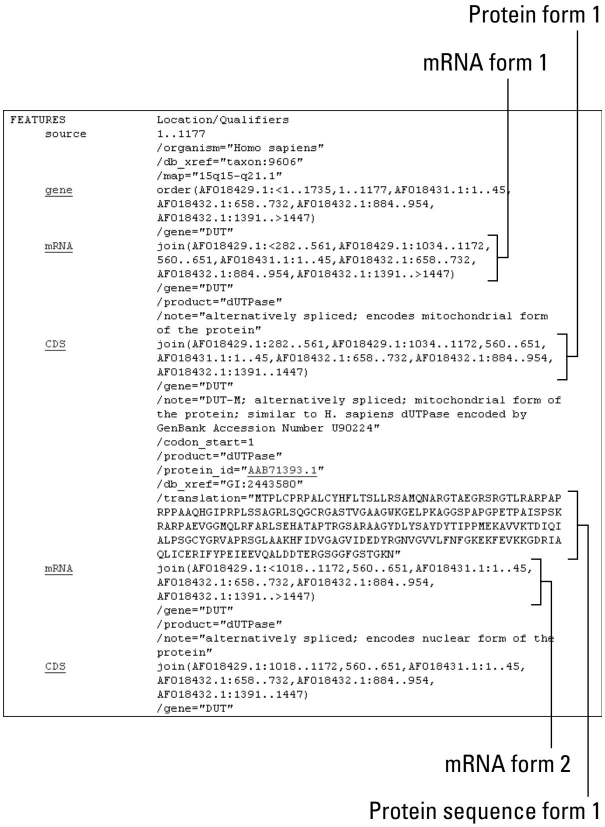
- The **gene** keyword introduces complex-looking formulas. Their purpose is to describe precisely the reconstruction of the various mRNAs spread over several separate entries.

  Concentrate on the first *gene order* formula to understand how this works:

  ```
  AF018429.1:<1..1735,1..1177,AF018431.1:1..45,
  AF018432.1:658..732,AF018432.1:884..954,
  AF018432.1:1391..>1447
  ```

  This is nothing more than an *exon splicing* recipe telling us how to reconstruct the mRNA sequence. All it says is:

  1. Take nucleotides from positions 1 to 1735 from entry AF018429.
  2. Add nucleotides from positions 1 to 1177 from the current entry (AF018430).
  3. Add nucleotides 1 to 45 from entry AF018431.
  4. Add nucleotides 658 to 732 from entry AF018432.



```
FEATURES             Location/Qualifiers
     source          1..1177
                     /organism="Homo sapiens"
                     /db_xref="taxon:9606"
                     /map="15q15-q21.1"
     gene            order(AF018429.1:<1..1735,1..1177,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
     mRNA            join(AF018429.1:<282..561,AF018429.1:1034..1172,
                     560..651,AF018431.1:1..45,AF018432.1:658..732,
                     AF018432.1:884..954,AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes mitochondrial form
                     of the protein"
     CDS             join(AF018429.1:282..561,AF018429.1:1034..1172,560..651,
                     AF018431.1:1..45,AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
                     /gene="DUT"
                     /note="DUT-M; alternatively spliced; mitochondrial form of
                     the protein; similar to H. sapiens dUTPase encoded by
                     GenBank Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71393.1"
                     /db_xref="GI:2443580"
                     /translation="MTPLCPRPALCYHFLTSLLRSAMQNARGTAEGRSRGTLRARPAP
                     RPPAAQHGIPRPLSSAGRLSQGCRGASTVGAAGWKGELPKAGGSPAPGPETPAISPSK
                     RARPAEVGGMQLRFARLSEHATAPTRGSARAAGYDLYSAYDYTIPPMEKAVVKTDIQI
                     ALPSGCYGRVAPRSGLAAKHFIDVGAGVIDEDYRGNVGVVLFNFGKEKFEVKKGDRIA
                     QLICERIFYPEIEEVQALDDTERGSGGFGSTGKN"
     mRNA            join(AF018429.1:<1018..1172,560..651,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..>1447)
                     /gene="DUT"
                     /product="dUTPase"
                     /note="alternatively spliced; encodes nuclear form of the
                     protein"
     CDS             join(AF018429.1:1018..1172,560..651,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
                     /gene="DUT"
```

Figure 3-8: GenBank entry AF018430: FEATURES table for a eukaryotic gene.

5. Add nucleotides 884 to 954 from the same entry (AF018432).
6. Add nucleotides 1391 to 1447 from the same entry again.

The < (less-than sign) at the beginning of the formula indicates that the gene might actually start before the indicated position, while the > (greater-than sign) at the end of the formula indicates that the gene might actually continue beyond the indicated position.

- **The mRNA:** The AF018430 entry has two mRNA fields, which are interesting because they illustrate the way GenBank represents alternative splicing patterns. Alternative splicing is a common property of higher eukaryotic gene expression. If you use the same parsing logic as gene order (which we describe in the preceding bullet), you can produce the results we summarize in Table 3-1.

Table 3-1 Mitochondrial or Nuclear dUTPase mRNAs in AF018430

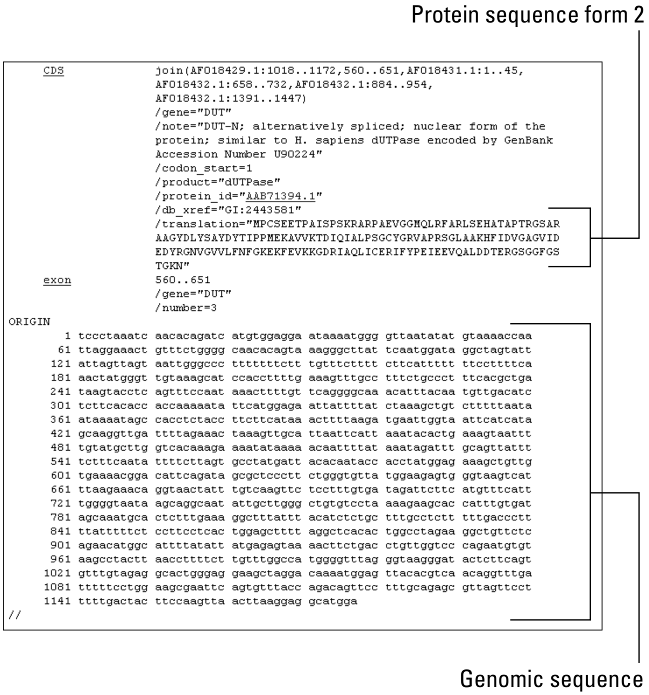
Mrna	*AF018429*	*AF018430*	*AF018431*	*AF018432*
Type 1 (Mitochondria)	282-561 1034-1172	560-651	1-45	658-732 884-954 1391-1447 ->
Type 2 (Nuclear)	<1018-1172	560-651	1-45	658-732 884-954 1391-1447 ->

Table 3-1 makes it easy to understand what's going on here:

- There are two alternative mRNA: type 1 (for the mitochondria) and type 2 (for the nucleus).
- The nuclear mRNA does not contain the first exon (stored in entry AF018429) although the mitochondrial mRNA uses this exon.
- The nuclear mRNA uses a part of the second exon (stored in the AF018429) in a different reading frame from its mitochondrial counterpart.
- The two protein sequences become identical on the third exon. You can see this in the two */translation* fields (*ETPAISPSK* and so on). See Figure 3-9 for the sequence of the nuclear form of the protein.

- The **exon** keyword indicates the position of the sole exon present in this sequence. This is near the end of the FEATURES part of the AF018430 entry (refer to Figure 3-9, positions 560–651). Look at AF018432 if you want to see multiple occurrences of the exon field in a single entry.

The sequence section, located in the bottom part of the entry, is formatted as usual. (Refer again to Figure 3-9.)



```
     CDS             join(AF018429.1:1018..1172,560..651,AF018431.1:1..45,
                     AF018432.1:658..732,AF018432.1:884..954,
                     AF018432.1:1391..1447)
                     /gene="DUT"
                     /note="DUT-N; alternatively spliced; nuclear form of the
                     protein; similar to H. sapiens dUTPase encoded by GenBank
                     Accession Number U90224"
                     /codon_start=1
                     /product="dUTPase"
                     /protein_id="AAB71394.1"
                     /db_xref="GI:2443581"
                     /translation="MPCSEETPAISPSKRARPAEVGGMQLRFARLSEHATAPTRGSAR
                     AAGYDLYSAYDYTIPPMEKAVVKTDIQIALPSGCYGRVAPRSGLAAKHFIDVGAGVID
                     EDYRGNVGVVLFNFGKEKFEVKKGDRIAQLICERIFYPEIEEVQALDDTERGSGGFGS
                     TGKN"
     exon            560..651
                     /gene="DUT"
                     /number=3
ORIGIN
        1 tccctaaatc aacacagatc atgtggagga ataaaatggg gttaatatat gtaaaaccaa
       61 ttaggaaact gtttctgggg caacacagta aagggcttat tcaatggata ggctagtatt
      121 attagttagt aattgggccc tttttttctt tgtttctttt cttcattttt ttccttttca
      181 aactatgggt tgtaaagcat ccaccttttg aaagtttgcc tttctgccct ttcacgctga
      241 taagtacctc agtttccaat aaacttttgt tcaggggcaa acatttacaa tgttgacatc
      301 tcttcacacc accaaaaata ttcatggaga attattttat ctaaagctgt ctttttaata
      361 ataaaatagc cacctctacc ttcttcataa acttttaaga tgaattggta attcatcata
      421 gcaaggttga ttttagaaac taaagttgca ttaattcatt aaatacactg aaagtaattt
      481 tgtatgcttg gtcacaaaga aaatataaaa acaattttat aaatagattt gcagttattt
      541 tctttcaata ttttcttagt gcctatgatt acacaatacc acctatggag aaagctgttg
      601 tgaaaacgga cattcagata gcgctccctt ctgggtgtta tggaagagtg ggtaagtcat
      661 ttaagaaaca ggtaactatt tgtcaagttc tcctttgtga tagattcttc atgtttcatt
      721 tggggtaata agcaggcaat attgcttggg ctgtgtccta aaagaagcac catttgtgat
      781 agcaaatgca ctctttgaaa ggctttattt acatctctgc tttgcctctt tttgaccctt
      841 ttatttttct ccttcctcac tggagctttt aggctcacac tggcctagaa ggctgttctc
      901 agaacatggc attttatatt atgagagtaa aacttctgac ctgttggtcc cagaatgtgt
      961 aagcctactt aaccttttct tgtttggcca tggggtttag ggtaagggat actcttcagt
     1021 gtttgtagag gcactgggag gaagctagga caaaatggag ttacacgtca acaggtttga
     1081 tttttcctgg aagcgaattc agtgtttacc agacagttcc tttgcagagc gttagttcct
     1141 ttttgactac ttccaagtta acttaaggag gcatgga
//
```

Figure 3-9: GenBank entry AF018430: Bottom and sequence part.

Understanding how to splice back the various nucleotide sequence fragments to form an mRNA and the associated coding regions from segmented GenBank entries was the main difficulty of this chapter. When you've done that, congratulations — you can sit back and relax while we continue our tour of GenBank.

Working with related GenBank entries

In every example that we've used so far, we knew the accession numbers for the entries we wanted to look up — X01714, U90223, or AF018430, for example. Normally you get these accession numbers by reading articles that explicitly mention them when reporting about the corresponding sequence.

That's a good strategy to start off with, but after you've accessed the first GenBank entry relevant to your work, you can retrieve other related genes by selecting the Related Sequences option in the pull-down menu that appears when clicking the Links link on the right side of the screen for each GenBank entry retrieved by its accession number (refer back to Figure 3-5). If you click this link for entry U90223, for example, it returns 37 human dUTPase-related GenBank entries. They include various mRNA forms and partial sequences, partial genomic sequences (around exons), and two large (154kb and 192kb) sequences of the 15q21.1 genomic region. There are some monkey sequences as well!

Retrieving GenBank entries without accession numbers

Although GenBank isn't the best database for keyword-based searches (see "Using a Gene-Centric Database," a bit later in this chapter, for a better method), querying GenBank by using gene or protein keywords rather than accession numbers is still possible. Suppose that you want to find the nucleotide sequence encoding the human dUTPase, but you don't have any handy accession numbers lying around. The following shows you how you should proceed, step by step:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   The NCBI PubMed home page appears.

2. **From the Search pull-down menu, choose Nucleotide.**

3. **Type** human [organism] AND dUTPase [Protein name] **in the Search window, and then click Go.**

   This search retrieves five GenBank entries: AH005568, AF018429, AF018430, AF018431, AF018432. They encompass exons 1 and 2 (AH005568), exon 3 (AF018430), exon 4 (AF018431), and exons 5, 6, and 7 (AF018432). The four last entries indicate the full amino-acid sequence of the two forms (nuclear and mitochondrial) of the dUTPase protein, as well as the alternative exon usage pattern. Not a bad start!

4. **Click the Links link to the right of the AF018432 entry ID line, and then choose Related Sequences from the pull-down menu that appears.**

   This retrieves a total of 20 entries. Among these entries, some contain mRNA sequences such as U90223.

   Note: We do not yet have all the entries related to human dUTPase in GenBank. By looking at some of the new entries we just pulled out, we realize that *dUTPase* is a nickname for "dUTP pyrophosphatase". Let's try to use these new terms in our next search:

5. **Type** human [organism] AND "dUTP pyrophosphatase" [Title] **(the "quotes" are important here) in the Search window, and then click the Go button.**

   Notice that we restricted the search to the `[Title]` field here — not as a protein name, which is what we did in Step 3.

This search now retrieves 20 GenBank entries that are not exactly the same ones we identified earlier! This illustrates the general difficulty in retrieving all entries relevant to a given subject, due to inconsistent usage of synonymous terms (such dUTPase, dUTP pyrophosphatase, or deoxyuridine triphosphatase) and fields. About half the current entries correspond to short, single-pass, partial mRNA sequences called *Expressed Sequence Tags* (ESTs).

To remove these ESTs from the list, you can use the Search-within-Limits trick we used with PubMed in Chapter 2:

1. **Open the Limits setting menu by clicking the Limits link (below the Search window).**
2. **Select the Exclude ESTs check box.**
3. **Scroll to the top of the form, and click the Go button.**

   Only eleven GenBank entries survive this particular limit-setting.

Feel free to use this handy Search-within-Limits protocol to try other field-restricted searches in GenBank.

Using a Gene-Centric Database

Besides the traditional GenBank, NCBI recently developed other databases (or database interfaces) that are more adapted to gene-centric queries.

Making a gene-centered query involves asking a question that relates directly to a specific gene, rather than going through all known pieces of sequences related to that gene. The main advantage of gene-centric databases is that they return results that are more synthetic than a long list of GenBank entries, and make much more sense to the biologists. Basically, you get the whole story at once.

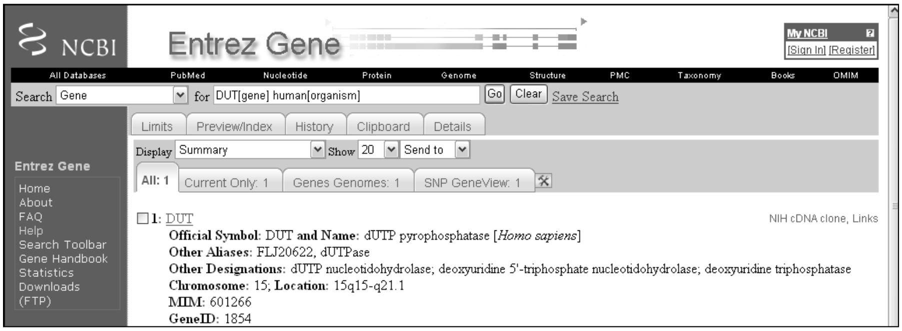
In this section, we take you through the Entrez/Gene resource available on the NCBI server. This resource makes it possible to gather important information related to a genetic *locus,* a specific place on a chromosome where a given gene has been identified. Thanks to this service, you can rapidly find out everything that is known about your favorite gene, or its genomic surrounding.

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   The NCBI/Entrez home page appears, eager to serve.
2. **From the Search pull-down menu, choose Gene.**
3. **Type** DUT [gene] human[organism] **in the For window, and then click Go.**

   You get a screen that looks like Figure 3-10.

Different queries such as `DUT[gene name] AND "homo sapiens"[organism]` produce the same result. More importantly, it's a way to land at the same place — by selecting Gene in the pull-down menu of Links associated with the nucleotide GenBank entry retrieved by accession number U90223 (see Figure 3-5). Now, that's a great way to link a piece of sequence with the corresponding gene!



Search Gene for DUT[gene] human[organism]

Display Summary Show 20

□ 1: DUT

Official Symbol: DUT **and Name:** dUTP pyrophosphatase [*Homo sapiens*]
Other Aliases: FLJ20622, dUTPase
Other Designations: dUTP nucleotidohydrolase; deoxyuridine 5'-triphosphate nucleotidohydrolase; deoxyuridine triphosphatase
Chromosome: 15; **Location:** 15q15-q21.1
MIM: 601266
GeneID: 1854

Figure 3-10: Entrez Gene entry page for human gene DUT.

The Results page you see in Figure 3-10 is your doorway to a wealth of information. By clicking the DUT link — or by changing the display option into Full Report — you can now get to a large body of information concerning this particular gene and its genomic environment.

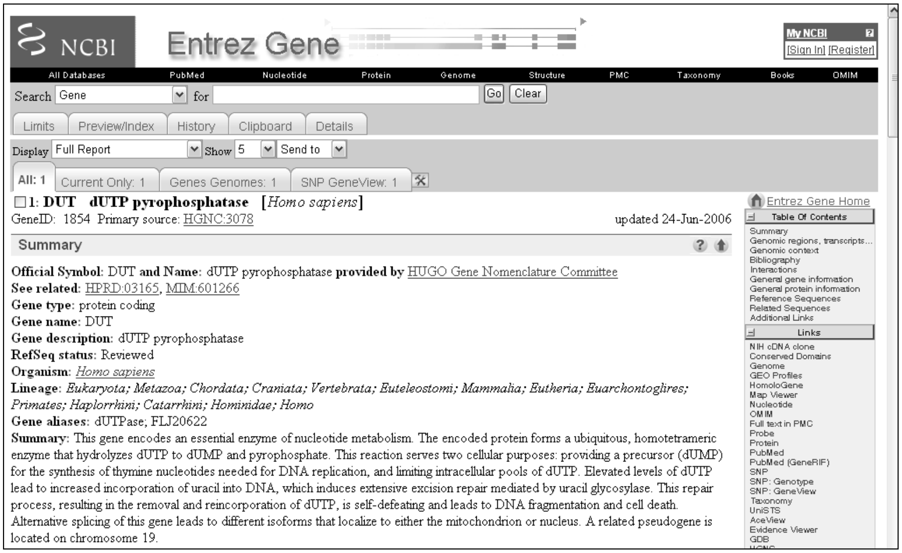
The top of the DUT entry (see Figure 3-11) provides a general description of what this gene is all about — and what function its products are known to perform, as well as a large variety of links (right-side menu) to other databases or NCBI files.

Figure 3-12 displays the next part of the entry: a schematic view of the Human DUT gene structure, with its seven exons used differentially and spread over 11,909 base pairs of genomic DNA on Chromosome 15.

The long entry then continues — additional sections provide information on potential interactions with other gene products, homologous sequences in other organisms, protein functions, and relevant metabolic pathways — as well as a list of all corresponding sequence entries in GenBank.

Each section, in turn, provides multiple links to help you get a complete (and sometimes redundant) picture of what's known about your gene. This one-stop shopping capacity illustrates the useful concept of a gene-centric database.

What you have here are all the types of mRNA sequences that have been observed and recorded in GenBank for this gene. You can see that variations mainly involve the two first exons. These variants (alternative transcripts) include the mitochondrial and nuclear forms of dUTPase (Table 3-1).



□1: **DUT dUTP pyrophosphatase** [*Homo sapiens*]

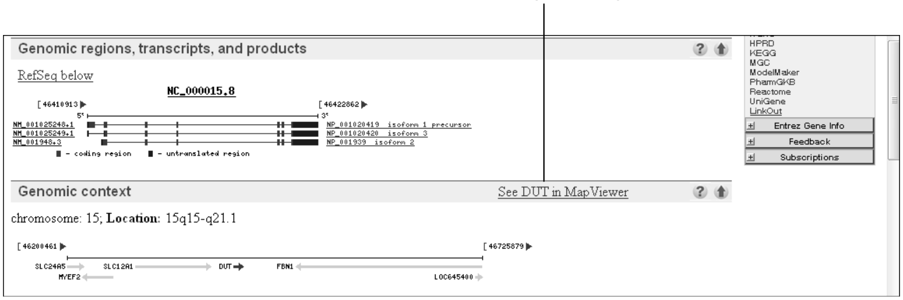
GeneID: 1854 Primary source: HGNC:3078 updated 24-Jun-2006

Summary

Official Symbol: DUT **and Name:** dUTP pyrophosphatase **provided by** HUGO Gene Nomenclature Committee
See related: HPRD:03165, MIM:601266
Gene type: protein coding
Gene name: DUT
Gene description: dUTP pyrophosphatase
RefSeq status: Reviewed
Organism: *Homo sapiens*
Lineage: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi; Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini; Catarrhini; Hominidae; Homo*
Gene aliases: dUTPase; FLJ20622
Summary: This gene encodes an essential enzyme of nucleotide metabolism. The encoded protein forms a ubiquitous, homotetrameric enzyme that hydrolyzes dUTP to dUMP and pyrophosphate. This reaction serves two cellular purposes: providing a precursor (dUMP) for the synthesis of thymine nucleotides needed for DNA replication, and limiting intracellular pools of dUTP. Elevated levels of dUTP lead to increased incorporation of uracil into DNA, which induces extensive excision repair mediated by uracil glycosylase. This repair process, resulting in the removal and reincorporation of dUTP, is self-defeating and leads to DNA fragmentation and cell death. Alternative splicing of this gene leads to different isoforms that localize to either the mitochondrion or nucleus. A related pseudogene is located on chromosome 19.

Figure 3-11: Top of the entry for human gene DUT.

Click here for a detailed map and experimental evidence.



chromosome: 15; **Location:** 15q15-q21.1

Figure 3-12: Genomic context and detailed alternative mRNAs for the Human DUT gene.

Click the See DUT in MapViewer link to go to a detailed display of the gene structure. You can review for yourself the experimental arguments in favor of the various mRNA models presented, right down to the nucleotide level.

Working with Whole-Genome Databases

The most recent genome-centric databases are the modern bioinformatic response to the proliferation of complete genome sequencing projects. The goal of these new types of resources is to gather all the information you need on a given organism, clearly separated from all the others, so you can more easily target your analyses on all genes from a specific genome. These new resources also promote comparisons of whole genomes with whole genomes, a new field of endeavor called *comparative genomics*.

These new databases still provide sequence information on isolated genes, but now bring in the additional capacity of showing them *in relation to each other* within the global framework of an entire genome.

The best way to interact with these databases is to zoom in and out until you reach the level of detail you need for the question you're interested in. Genome-centric databases also provide the easiest way to gather every gene/protein sequence of a given organism at once, with only a few mouse clicks.

We start our exploration of whole genome databases by taking a peek at the Viral Genome section available on the NCBI server.

Working with complete viral genomes

Viruses are fascinating objects, on the edge of the living world. They function as minimal molecular bits of machinery, cleverly designed to ensure the multiplication of nucleic-acid molecules (the *viral genome*) at the expense of cellular hosts (eukaryotic, bacterial, or archaebacterial). While going about their business, viruses might go unnoticed — or trigger dreadful diseases and epidemics, such as smallpox, poliomyelitis, or AIDS. Viral genomes come in all varieties of biochemical forms (RNA/DNA, circular/linear, single/double stranded) and size (a few kb to a million bp).

Visiting the NCBI Viral Genome resource is a great way to find out more about them. To demonstrate, we show you what you can find out about the dreadful AIDS-causing, type-1 human immunodeficiency virus, or HIV-1.

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

   You've probably memorized this address by now. The NCBI PubMed home page appears.

2. **On the black menu bar at the top of the form, click Genome.**

   This takes you to the Entrez Genome page.

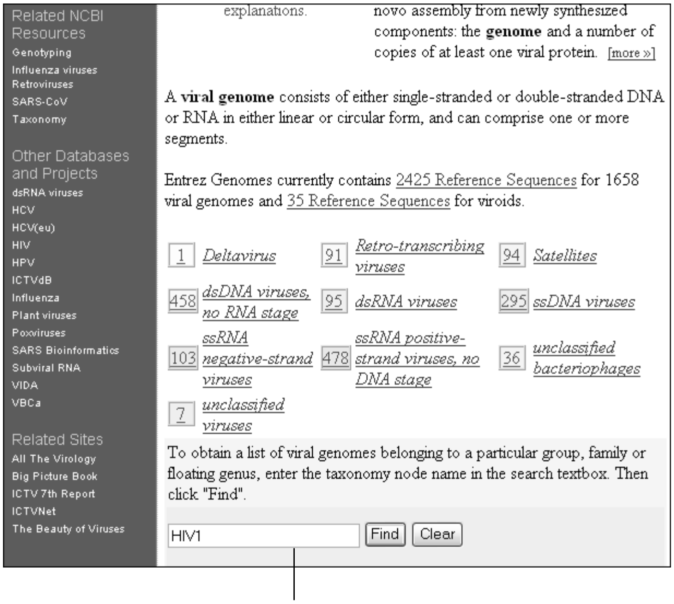
3. **Click the Viruses link (on the right side of the form).**

   The Viral Genomes reference page appears.

4. **Scroll down the Viral Reference Genomes page until you reach the table of available viral-genome sequences grouped by class (Deltavirus, Retroid viruses, and so on), as shown in Figure 3-13.**

5. **Type** HIV1 **in the Search window, and then click the Find button.**

   Your browser returns a nice global summary of the HIV-1 genome, as shown in Figure 3-14. At the bottom, a clickable picture indicates the identity and respective positions of all the genes.



explanations. novo assembly from newly synthesized components: the **genome** and a number of copies of at least one viral protein. [more »]

A **viral genome** consists of either single-stranded or double-stranded DNA or RNA in either linear or circular form, and can comprise one or more segments.

Entrez Genomes currently contains 2425 Reference Sequences for 1658 viral genomes and 35 Reference Sequences for viroids.

1	*Deltavirus*	91	*Retro-transcribing viruses*	94	*Satellites*
458	*dsDNA viruses, no RNA stage*	95	*dsRNA viruses*	295	*ssDNA viruses*
103	*ssRNA negative-strand viruses*	478	*ssRNA positive-strand viruses, no DNA stage*	36	*unclassified bacteriophages*
7	*unclassified viruses*				

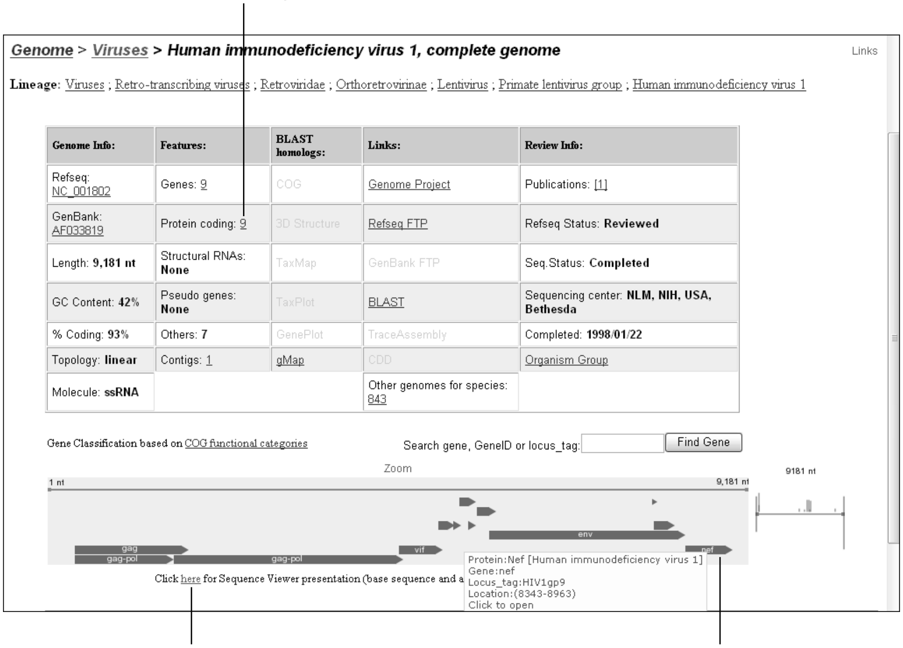
To obtain a list of viral genomes belonging to a particular group, family or floating genus, enter the taxonomy node name in the search textbox. Then click "Find".

HIV1 [Find] [Clear]

Enter the name of your virus of interest.

Figure 3-13: The table of all available viral genome sequences.

Click here for all proteins.



***Genome* > *Viruses* > Human immunodeficiency virus 1, complete genome**

Links

Lineage: Viruses ; Retro-transcribing viruses ; Retroviridae ; Orthoretrovirinae ; Lentivirus ; Primate lentivirus group ; Human immunodeficiency virus 1

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_001802	Genes: 9	COG	Genome Project	Publications: [1]
GenBank: AF033819	Protein coding: 9	3D Structure	Refseq FTP	Refseq Status: **Reviewed**
Length: **9,181 nt**	Structural RNAs: **None**	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **42%**	Pseudo genes: **None**	TaxPlot	BLAST	Sequencing center: **NLM, NIH, USA, Bethesda**
% Coding: **93%**	Others: **7**	GenePlot	TraceAssembly	Completed: **1998/01/22**
Topology: **linear**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **ssRNA**			Other genomes for species: 843	

Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag: [Find Gene]

Protein:Nef [Human immunodeficiency virus 1]
Gene:nef
Locus_tag:HIV1gp9
Location:(8343-8963)
Click to open

Click here for Sequence Viewer presentation (base sequence and a

Click here for a live map.

Click here to get gene details.

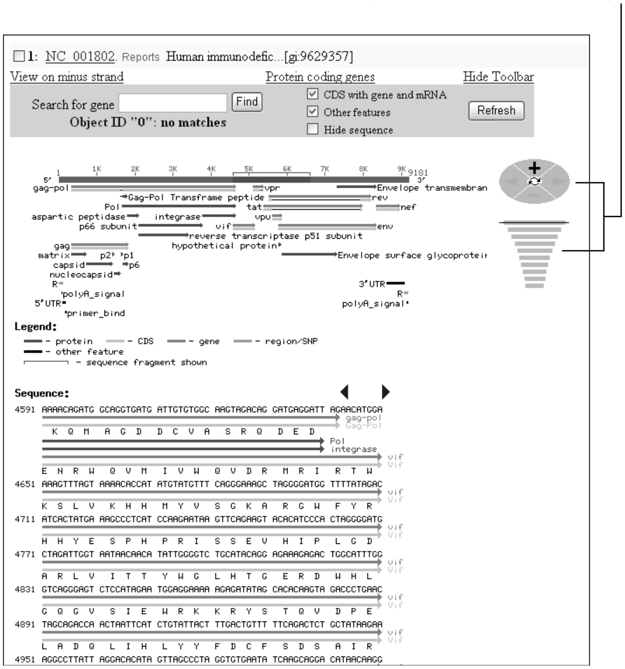
Figure 3-14: General structure of the HIV-1 genome.

Similar pages are available for all viruses, regardless of their sizes.

TIP

Clicking the here link (bottom of Figure 3-14) gets you a live map, allowing you to zoom in on any genome region, down to the nucleotide sequence level, as shown in Figure 3-15.

Navigate/zoom in on the viral genome using these buttons.



□ 1: NC_001802. Reports Human immunodefic...[gi:9629357]

View on minus strand　　Protein coding genes　　Hide Toolbar

Search for gene [ ] Find

Object ID "0": no matches

☑ CDS with gene and mRNA
☑ Other features
□ Hide sequence

Refresh

Legend:

- protein
- CDS
- gene
- region/SNP
- other feature
- sequence fragment shown

Sequence:

```
4591 AAAACAGATG GCAGGTGATG ATTGTGTGGC AAGTAGACAG GATGAGGATT AGAACATGGA
                                                              gag-pol
                                                              Gag-Pol
       K  Q  M   A  G  D   D  C  V  A   S  R  Q   D  E  D
                                                           Pol
                                                           integrase
                                                                    vif
                                                                    Vif
     E  N  R  W   Q  V  M   I  V  W   Q  V  D  R   M  R  I   R  T  W
4651 AAAGTTTAGT AAAACACCAT ATGTATGTTT CAGGGAAAGC TAGGGGATGG TTTTATAGAC
                                                                    vif
                                                                    Vif
     K  S  L  V   K  H  H   M  Y  V   S  G  K  A   R  G  W   F  Y  R
4711 ATCACTATGA AAGCCCTCAT CCAAGAATAA GTTCAGAAGT ACACATCCCA CTAGGGGATG
                                                                    vif
                                                                    Vif
     H  H  Y  E   S  P  H   P  R  I   S  S  E  V   H  I  P   L  G  D
4771 CTAGATTGGT AATAACAACA TATTGGGGTC TGCATACAGG AGAAAGAGAC TGGCATTTGG
                                                                    vif
                                                                    Vif
     A  R  L  V   I  T  T   Y  W  G   L  H  T  G   E  R  D   W  H  L
4831 GTCAGGGAGT CTCCATAGAA TGGAGGAAAA AGAGATATAG CACACAAGTA GACCCTGAAC
                                                                    vif
                                                                    Vif
     G  Q  G  V   S  I  E   W  R  K   K  R  Y  S   T  Q  V   D  P  E
4891 TAGCAGACCA ACTAATTCAT CTGTATTACT TTGACTGTTT TTCAGACTCT GCTATAAGAA
                                                                    vif
                                                                    Vif
     L  A  D  Q   L  I  H   L  Y  Y   F  D  C  F   S  D  S   A  I  R
4951 AGGCCTTATT AGGACACATA GTTAGCCCTA GGTGTGAATA TCAAGCAGGA CATAACAAGG
```

Figure 3-15: Zooming in on the HIV-1 genome: the Pol/Vif overlap.

Figure 3-15 shows a position in the genome where the end of the Pol gene slightly overlaps with the beginning of the Vif gene. Viruses commonly have the same nucleotide sequence involved in the making of two different amino-acid sequences.

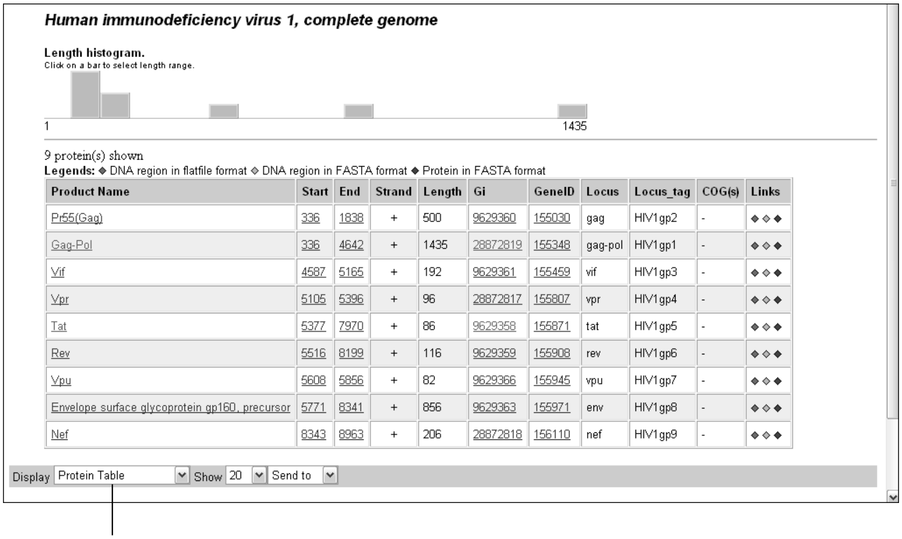
6. **Click your browser's Back button.**

   This takes you back to the HIV-1 Genome entry page. (Refer to Figure 3-14.)

7. **Click the number (9) following Protein coding (second column and row of the table).**

   The Protein List page appears, as shown in Figure 3-16. Here, you can retrieve the DNA and protein sequences in different formats — GenBank format, or FASTA format — by clicking one of the three lozenges in the rightmost column.

If you want, you can download the complete set of protein sequences by selecting Protein FASTA in the Display pull-down menu.



Human immunodeficiency virus 1, complete genome

Length histogram.
Click on a bar to select length range.

9 protein(s) shown
Legends: ◆ DNA region in flatfile format ◇ DNA region in FASTA format ◆ Protein in FASTA format

Product Name	Start	End	Strand	Length	Gi	GeneID	Locus	Locus_tag	COG(s)	Links
Pr55(Gag)	336	1838	+	500	9629360	155030	gag	HIV1gp2	-	◆ ◇ ◆
Gag-Pol	336	4642	+	1435	28872819	155348	gag-pol	HIV1gp1	-	◆ ◇ ◆
Vif	4587	5165	+	192	9629361	155459	vif	HIV1gp3	-	◆ ◇ ◆
Vpr	5105	5396	+	96	28872817	155807	vpr	HIV1gp4	-	◆ ◇ ◆
Tat	5377	7970	+	86	9629358	155871	tat	HIV1gp5	-	◆ ◇ ◆
Rev	5516	8199	+	116	9629359	155908	rev	HIV1gp6	-	◆ ◇ ◆
Vpu	5608	5856	+	82	9629366	155945	vpu	HIV1gp7	-	◆ ◇ ◆
Envelope surface glycoprotein gp160, precursor	5771	8341	+	856	9629363	155971	env	HIV1gp8	-	◆ ◇ ◆
Nef	8343	8963	+	206	28872818	156110	nef	HIV1gp9	-	◆ ◇ ◆

Display: Protein Table   Show: 20   Send to

Figure 3-16: Downloading HIV-1 gene and protein sequences.

Select Protein FASTA here to download all sequences.

Working with complete bacterial genomes

NCBI Entrez offers a nice interface to all publicly available complete bacterial genome sequences. For a quick tour, do the following:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/entrez/`**.**

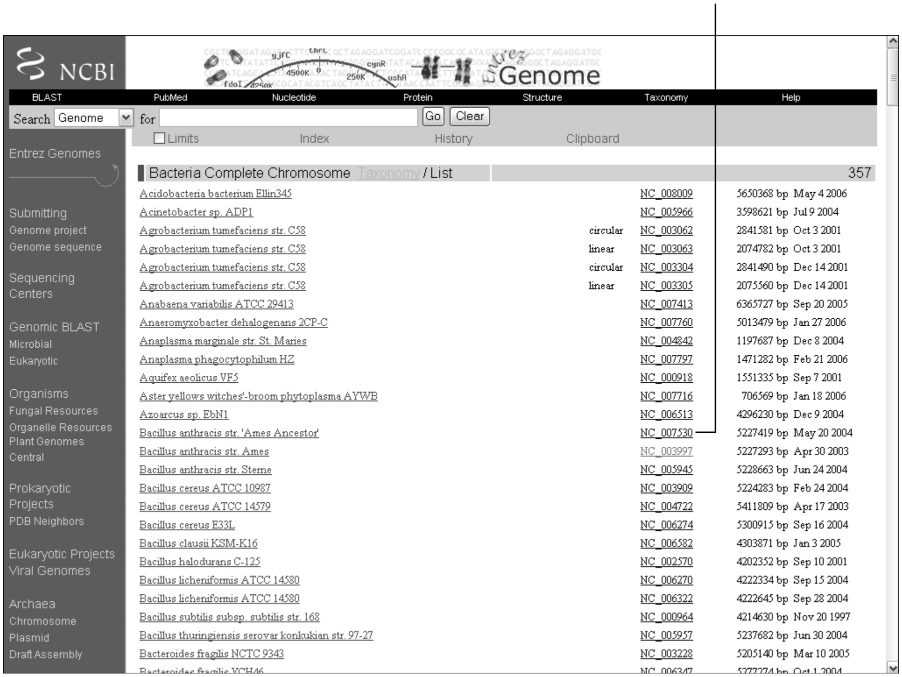
   The NCBI PubMed home page dutifully appears.

2. **Click Genome on the black menu bar near the top of the form.**

   This takes you to the Entrez Genome page.

3. **In the left dark-blue margin, click the Chromosome link located under the Bacteria heading.**

   This step returns a listing of all bacteria whose chromosomes have been fully sequenced. (Directly clicking Bacteria would have given you a longer list, including parasitic DNA segment called plasmids.) Figure 3-17 shows the top of the list.



Click here.

Bacteria Complete Chromosome Taxonomy / List 357

Acidobacteria bacterium Ellin345		NC_008009	5650368 bp	May 4 2006
Acinetobacter sp. ADP1		NC_005966	3598621 bp	Jul 9 2004
Agrobacterium tumefaciens str. C58	circular	NC_003062	2841581 bp	Oct 3 2001
Agrobacterium tumefaciens str. C58	linear	NC_003063	2074782 bp	Oct 3 2001
Agrobacterium tumefaciens str. C58	circular	NC_003304	2841490 bp	Dec 14 2001
Agrobacterium tumefaciens str. C58	linear	NC_003305	2075560 bp	Dec 14 2001
Anabaena variabilis ATCC 29413		NC_007413	6365727 bp	Sep 20 2005
Anaeromyxobacter dehalogenans 2CP-C		NC_007760	5013479 bp	Jan 27 2006
Anaplasma marginale str. St. Maries		NC_004842	1197687 bp	Dec 8 2004
Anaplasma phagocytophilum HZ		NC_007797	1471282 bp	Feb 21 2006
Aquifex aeolicus VF5		NC_000918	1551335 bp	Sep 7 2001
Aster yellows witches'-broom phytoplasma AYWB		NC_007716	706569 bp	Jan 18 2006
Azoarcus sp. EbN1		NC_006513	4296230 bp	Dec 9 2004
Bacillus anthracis str. 'Ames Ancestor'		NC_007530	5227419 bp	May 20 2004
Bacillus anthracis str. Ames		NC_003997	5227293 bp	Apr 30 2003
Bacillus anthracis str. Sterne		NC_005945	5228663 bp	Jun 24 2004
Bacillus cereus ATCC 10987		NC_003909	5224283 bp	Feb 24 2004
Bacillus cereus ATCC 14579		NC_004722	5411809 bp	Apr 17 2003
Bacillus cereus E33L		NC_006274	5300915 bp	Sep 16 2004
Bacillus clausii KSM-K16		NC_006582	4303871 bp	Jan 3 2005
Bacillus halodurans C-125		NC_002570	4202352 bp	Sep 10 2001
Bacillus licheniformis ATCC 14580		NC_006270	4222334 bp	Sep 15 2004
Bacillus licheniformis ATCC 14580		NC_006322	4222645 bp	Sep 28 2004
Bacillus subtilis subsp. subtilis str. 168		NC_000964	4214630 bp	Nov 20 1997
Bacillus thuringiensis serovar konkukian str. 97-27		NC_005957	5237682 bp	Jun 30 2004
Bacteroides fragilis NCTC 9343		NC_003228	5205140 bp	Mar 10 2005
Bacteroides fragilis YCH46		NC_006347	5277274 bp	Oct 1 2004

Figure 3-17: The list of all available bacterial genome sequences (top part).

4. **Click the NC_007530 link corresponding to the dangerous bioterrorism agent *Bacillus anthracis*.**

   Your browser displays a table that summarizes the content and features of the bacterial genome. It's similar to what we already got for viruses (see Figure 3-18).

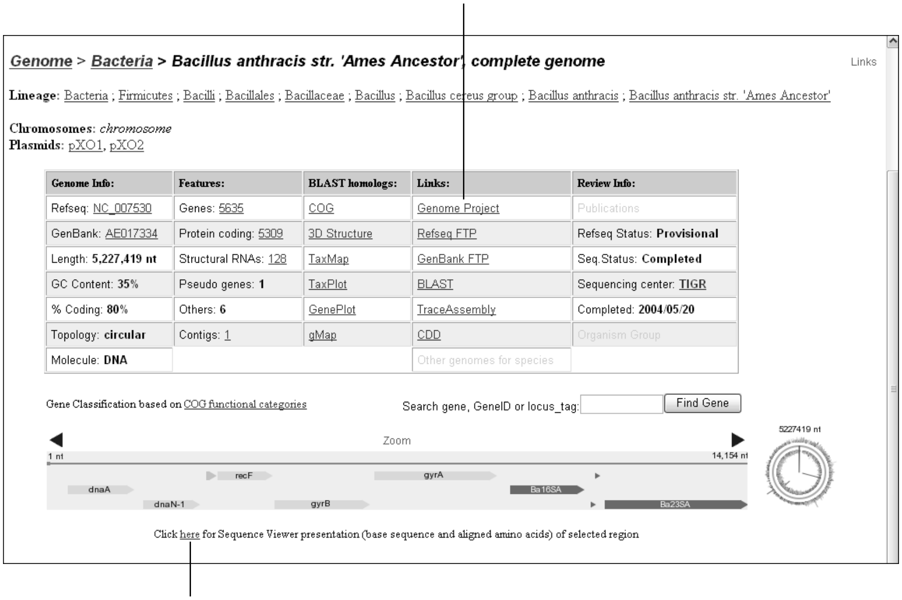
   TIP

   Clicking the Here link near the bottom gets you the same type of live map we saw with the HIV-1 virus (see the previous section), although the genome is now much larger. On this map, you can click to zoom into a particular region of the genome. The genome summary table contains numerous links to analyses pre-computed for all the genes (function, similarity, evolutionary relationships) that we leave you to try out for yourself.

5. **Back in the Summary table (refer to Figure 3-18) click the Genome Project link.**

   Doing so gets you a quick description of the bacterium, along with a picture, details about its habitat, the disease it causes, and so on — as well as the reasons why it was important to decipher its genome sequence in the first place. (See Figure 3-19.)

Click here to learn about the bacteria.



Genome > Bacteria > Bacillus anthracis str. 'Ames Ancestor', complete genome

Lineage: Bacteria ; Firmicutes ; Bacilli ; Bacillales ; Bacillaceae ; Bacillus ; Bacillus cereus group ; Bacillus anthracis ; Bacillus anthracis str. 'Ames Ancestor'

Chromosomes: *chromosome*
Plasmids: pXO1, pXO2

Genome Info:	Features:	BLAST homologs:	Links:	Review Info:
Refseq: NC_007530	Genes: 5635	COG	Genome Project	Publications
GenBank: AE017334	Protein coding: 5309	3D Structure	Refseq FTP	Refseq Status: **Provisional**
Length: **5,227,419 nt**	Structural RNAs: 128	TaxMap	GenBank FTP	Seq.Status: **Completed**
GC Content: **35%**	Pseudo genes: **1**	TaxPlot	BLAST	Sequencing center: **TIGR**
% Coding: **80%**	Others: **6**	GenePlot	TraceAssembly	Completed: **2004/05/20**
Topology: **circular**	Contigs: 1	gMap	CDD	Organism Group
Molecule: **DNA**			Other genomes for species	

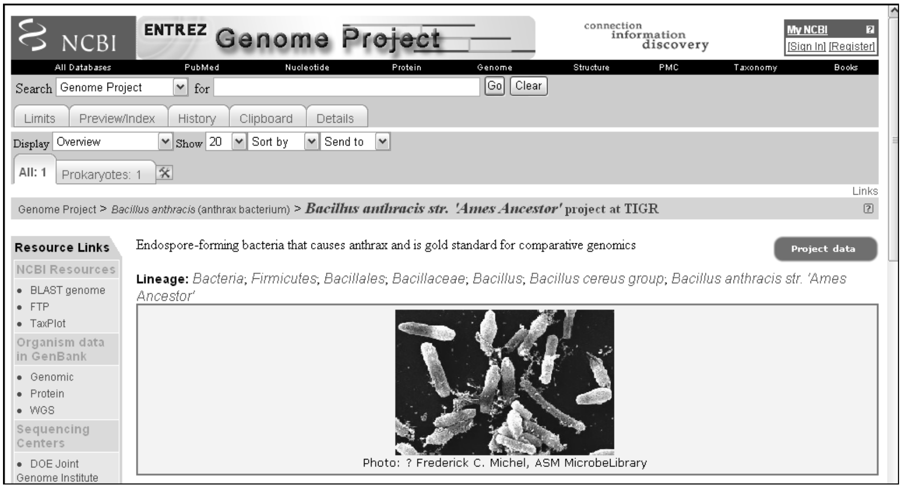
Gene Classification based on COG functional categories

Search gene, GeneID or locus_tag:

Click here for Sequence Viewer presentation (base sequence and aligned amino acids) of selected region

Click here for a live map.

Figure 3-18: *Bacillus anthracis* (strain Ames ancestor) genome summary.



Search Genome Project for

Display Overview Show 20 Sort by Send to

Genome Project > *Bacillus anthracis* (anthrax bacterium) > ***Bacillus anthracis str. 'Ames Ancestor'*** project at TIGR

Resource Links

NCBI Resources
- BLAST genome
- FTP
- TaxPlot

Organism data in GenBank
- Genomic
- Protein
- WGS

Sequencing Centers
- DOE Joint Genome Institute

Endospore-forming bacteria that causes anthrax and is gold standard for comparative genomics

Lineage: *Bacteria; Firmicutes; Bacillales; Bacillaceae; Bacillus; Bacillus cereus group; Bacillus anthracis str. 'Ames Ancestor'*

Photo: ? Frederick C. Michel, ASM MicrobeLibrary

Figure 3-19: Genome Project page for *Bacillus anthracis* (strain Ames ancestor).

More bacterial genomics at TIGR

The Institute for Genome Research (better known by its acronym TIGR) is home to a team of scientists who pioneered the field of bacterial genomics. In 1995, these scientists produced the two first complete bacterial genomes:

Hemophilus influenzae and *Mycoplasma genitalium*. Since then, they have contributed to more than 70 complete bacterial genomes, with more on the way.

TIGR offers a site that is quite complementary to the NCBI resource because it keeps track of all ongoing bacterial genome sequencing projects (not only of the completed ones), propose its own variety of well-integrated analysis tools for use, and offers the possibility to run similarity searches on its genome sequence data well before the actual completion of the projects themselves. In addition, TIGR's scientists offer a specific perspective on the genomes they have sequenced through their own genome viewing software.

To pay them a quick visit, do the following:

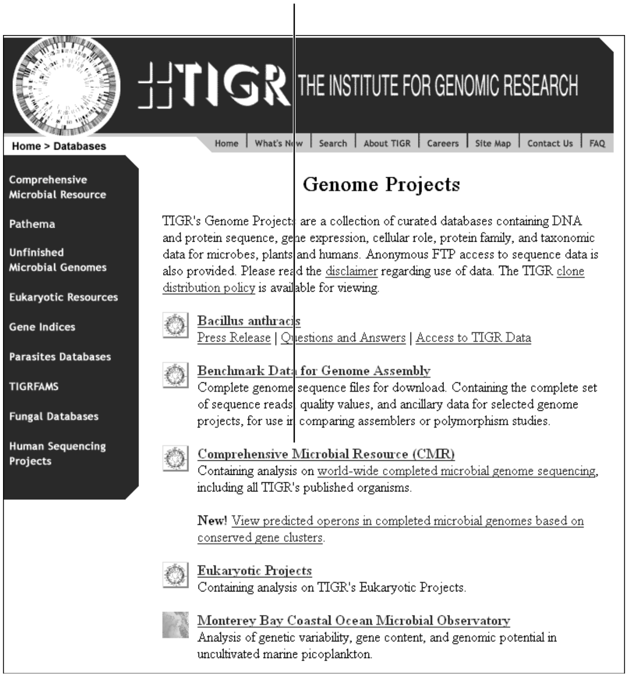
1. **Point your browser to** `www.tigr.org/tdb/`**.**

   The TIGR home page appears, as shown in Figure 3-20.

2. **Click the** Comprehensive Microbial Resource (CMR) **link.**

   A long list of micro-organisms appears, from which you get to select the one you are interested in.

3. **Select one microbe in the list, and then explore the various analyses tools by clicking the** Genome Searches**,** Genome Toolbox**, and** Genome Analyses **links at the top of each microbe-specific page.**


TIGR THE INSTITUTE FOR GENOMIC RESEARCH
Home > Databases
Home | What's New | Search | About TIGR | Careers | Site Map | Contact Us | FAQ
Comprehensive Microbial Resource
Pathema
Unfinished Microbial Genomes
Eukaryotic Resources
Gene Indices
Parasites Databases
TIGRFAMS
Fungal Databases
Human Sequencing Projects

Click here.

Genome Projects

TIGR's Genome Projects are a collection of curated databases containing DNA and protein sequence, gene expression, cellular role, protein family, and taxonomic data for microbes, plants and humans. Anonymous FTP access to sequence data is also provided. Please read the disclaimer regarding use of data. The TIGR clone distribution policy is available for viewing.

Bacillus anthracis
Press Release | Questions and Answers | Access to TIGR Data

Benchmark Data for Genome Assembly
Complete genome sequence files for download. Containing the complete set of sequence reads, quality values, and ancillary data for selected genome projects, for use in comparing assemblers or polymorphism studies.

Comprehensive Microbial Resource (CMR)
Containing analysis on world-wide completed microbial genome sequencing, including all TIGR's published organisms.

New! View predicted operons in completed microbial genomes based on conserved gene clusters.

Eukaryotic Projects
Containing analysis on TIGR's Eukaryotic Projects.

Monterey Bay Coastal Ocean Microbial Observatory
Analysis of genetic variability, gene content, and genomic potential in uncultivated marine picoplankton.

Figure 3-20: The Genome Projects site at The Institute for Genome Research (TIGR).

Microbes from the environment at DoE

The U.S. Department of Energy (DoE) is also a main player in microbial genomics. Its Joint Genome Institute specializes in the study of organisms that are either (a) important for preserving (or de-polluting) our environment, or (b) offering some new perspective in solving the incoming worldwide energy crisis (such as cheap ways of producing hydrogen).

To take a look at its data, follow these steps:

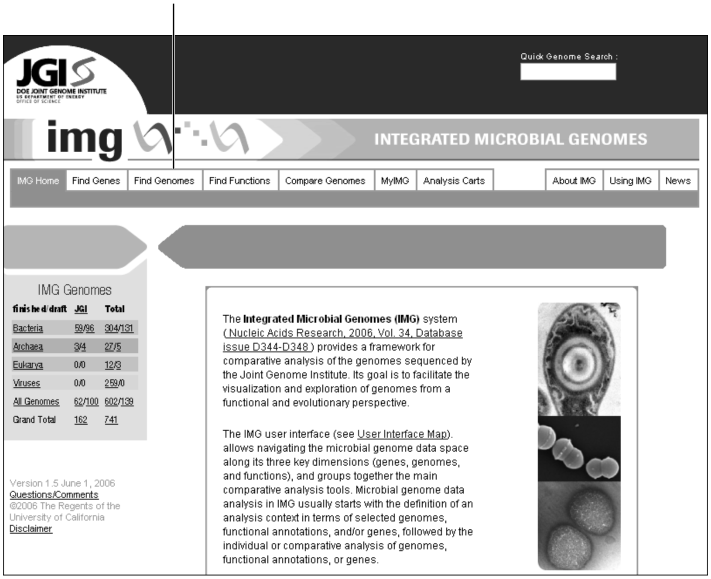
1. **Point your browser to** `img.jgi.doe.gov/`**.**

   The Integrated Microbial Genomes resources home page appears, as shown in Figure 3-21. As was the case with the NCBI and TIGR sites, specific sets of tools are proposed. The next step looks at the IMG genome display — one of the more useful IMG resources.

2. **To access the IMG Genome display, first click the Find Genomes link.**

   A table of the available organisms appears.

Click here to start.



IMG Genomes

finished/draft	JGI	Total
Bacteria	59/96	304/131
Archaea	3/4	27/5
Eukarya	0/0	12/3
Viruses	0/0	259/0
All Genomes	62/100	602/139
Grand Total	162	741

Version 1.5 June 1, 2006
Questions/Comments
©2006 The Regents of the University of California
Disclaimer

The **Integrated Microbial Genomes (IMG)** system (Nucleic Acids Research, 2006, Vol. 34, Database issue D344-D348) provides a framework for comparative analysis of the genomes sequenced by the Joint Genome Institute. Its goal is to facilitate the visualization and exploration of genomes from a functional and evolutionary perspective.

The IMG user interface (see User Interface Map). allows navigating the microbial genome data space along its three key dimensions (genes, genomes, and functions), and groups together the main comparative analysis tools. Microbial genome data analysis in IMG usually starts with the definition of an analysis context in terms of selected genomes, functional annotations, and/or genes, followed by the individual or comparative analysis of genomes, functional annotations, or genes.

Figure 3-21: The Integrated Microbial Genomes site at the DoE Joint Genome Institute.

3. **Select an available organism by checking the corresponding box, then clicking Save Selections.**

   The first organism, *Aeropyrum pernix,* would be a good choice for now.

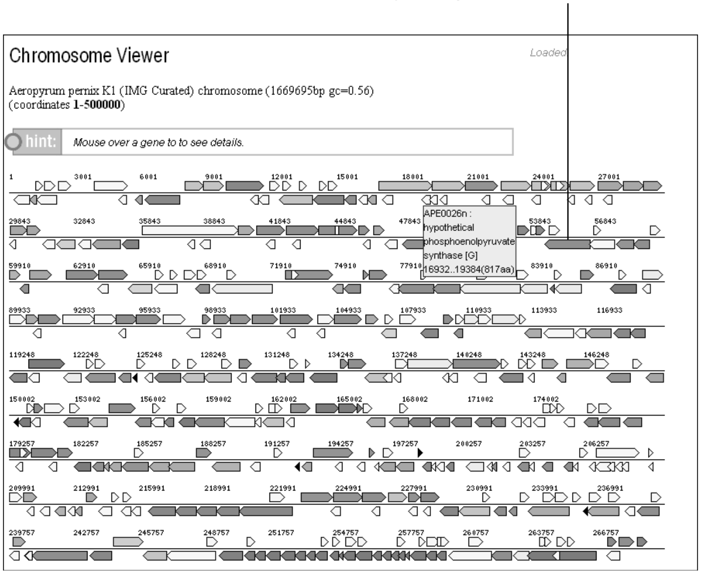
4. **In the new page that appears, confirm your selection by clicking Compare Genomes.**

   A Genome Statistics page appears.

5. **In the Genome Statistics page, click the number one in the DNA scaffolds row.**

6. **In the new page that appears, select a coordinate range (for instance 1 . . . 500000).**

   This results in a live display of the microbe gene content in that range, as shown in Figure 3-22. Putting the mouse pointer over the gene symbol gets you its name — and clicking the symbol provides you with a wealth of information on the corresponding protein.



Click on a gene symbol to find out more about its function.

Chromosome Viewer

Loaded

Aeropyrum pernix K1 (IMG Curated) chromosome (1669695bp gc=0.56)
(coordinates **1-500000**)

hint: *Mouse over a gene to to see details.*

APE0026n :
hypothetical
phosphoenolpyruvate
synthase [G]
16932..19384(817aa)

Figure 3-22: The chromosome viewer at the Integrated Microbial Genomes DoE resource.

Exploring the Human Genome

The sequencing of the human genome is probably one of the greatest scientific accomplishments of modern times. With 3000 million (and then some) nucleotides spread over 23 chromosomes, this genome is definitely a complex object to deal with. The major task ahead for bioinformatics is to integrate all the past, present, and future information that human genes contain in a maintainable, user-friendly resource.

If you want to make sense of human data, you must have clear ideas on the current state of the data:

- The complete nucleotide sequence of the human genome is now at hand — except for a few thousand holes that are still being filled up. This is a reason why new versions (also called "assemblies") of the — almost-finished — human genome are periodically released. This state of flux is true for all large animal genomes.
- This sequence was obtained in raw format; the next challenge is the annotation of the raw data — creating a detailed and accurate FEATURES table of the human genome.
- Throughout the world, new information is generated daily on human gene properties and functions, using a wide array of techniques. Ideally, someone has to gather it all, package it nicely, and offer it to the whole research community for free!

Having doubts that the last goal will ever be attainable in this less-than-perfect world? Well, think again, because this is where we're taking you right now.

Finding out about the Ensembl project

The Internet home page of Ensembl (`www.ensembl.org`) says it all: Ensembl is a joint project between the European Bioinformatics Institute (EBI) and the Sanger Institute, both located near Cambridge (U.K.). Together they've developed an integrated database and software system to produce and maintain automatic annotations for the genomes of animals, with a special attention to our closest relatives: the vertebrates. This project — like the others we describe in this chapter — relies heavily on the collaboration of a large number of individual laboratories. It all started as part of the International Human Genome Project, continued for the Mouse Genome project, and is now being pursued for other animals. Data and software are also freely flowing among numerous national database and bioinformatics centers from all over the world, allowing a complex cross-linking to take place.

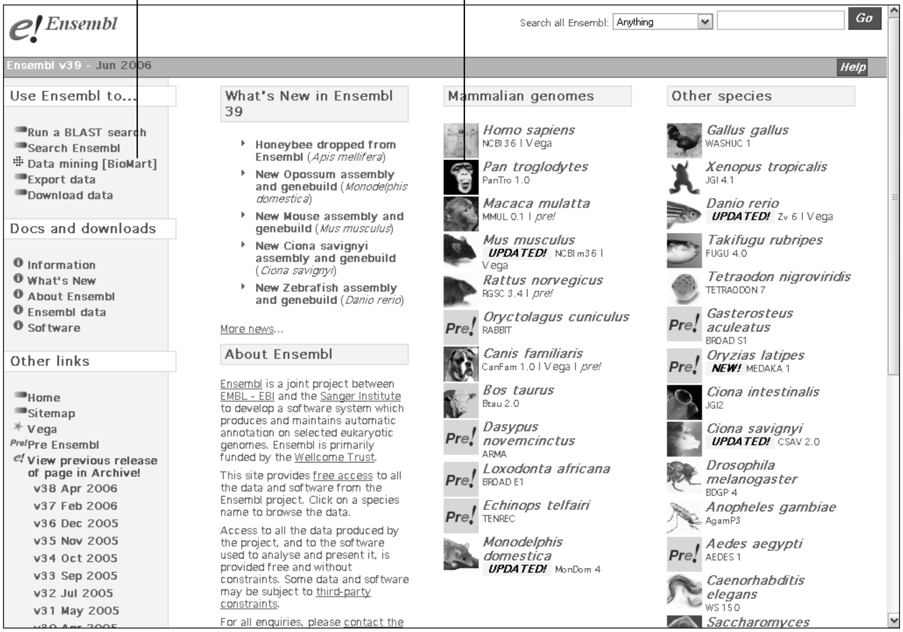
Getting started on the Ensembl site

Considering how complex the human genome is, you may not be surprised to find that you can attack the Ensembl resources from many different angles. To quickly find out about your options, the best thing to do is to jump on the guided tour that Ensembl proposes on its home page. Here's how it's done:

1. **Point your browser to `www.ensembl.org/`.**

   The impressive Ensembl home page appears, as shown in Figure 3-23.

The page densely filled up with links, much too numerous to be explored in detail in this chapter. You can literally spend weeks navigating all the possibilities. For now, let's focus on the right half of the page, where links to specific animal genome areas are indicated by small pictures. Thanks to them, you already learned the Latin species name for chimpanzee (*Pan troglodytes!*), rabbit (*Oryctolagus cunicumus!*), or armadillo (*Dasypus novemcinctus!*). However, the main reason we're here is to take a look at the human genome, so let's click the DaVinci-inspired *Homo sapiens* icon.



Use BioMart for serious data mining.

Click here to start navigating the human genome.

e! Ensembl

Search all Ensembl: Anything [Go]

Ensembl v39 - Jun 2006 [Help]

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Other links

- Home
- Sitemap
- Vega
- Pre Ensembl
- View previous release of page in Archive!
  - v38 Apr 2006
  - v37 Feb 2006
  - v36 Dec 2005
  - v35 Nov 2005
  - v34 Oct 2005
  - v33 Sep 2005
  - v32 Jul 2005
  - v31 May 2005

What's New in Ensembl 39

- **Honeybee dropped from Ensembl** (*Apis mellifera*)
- **New Opossum assembly and genebuild** (*Monodelphis domestica*)
- **New Mouse assembly and genebuild** (*Mus musculus*)
- **New Ciona savignyi assembly and genebuild** (*Ciona savignyi*)
- **New Zebrafish assembly and genebuild** (*Danio rerio*)

More news...

About Ensembl

Ensembl is a joint project between EMBL - EBI and the Sanger Institute to develop a software system which produces and maintains automatic annotation on selected eukaryotic genomes. Ensembl is primarily funded by the Wellcome Trust.

This site provides free access to all the data and software from the Ensembl project. Click on a species name to browse the data.

Access to all the data produced by the project, and to the software used to analyse and present it, is provided free and without constraints. Some data and software may be subject to third-party constraints.

For all enquiries, please contact the

Mammalian genomes

Homo sapiens
NCBI 36 | Vega

Pan troglodytes
PanTro 1.0

Macaca mulatta
MMUL 0.1 | *pre!*

Mus musculus
UPDATED! NCBI m36 | Vega

Rattus norvegicus
RGSC 3.4 | *pre!*

Pre! *Oryctolagus cuniculus*
RABBIT

Canis familiaris
CanFam 1.0 | Vega | *pre!*

Bos taurus
Btau 2.0

Pre! *Dasypus novemcinctus*
ARMA

Pre! *Loxodonta africana*
BROAD E1

Pre! *Echinops telfairi*
TENREC

Monodelphis domestica
UPDATED! MonDom 4

Other species

Gallus gallus
WASHUC 1

Xenopus tropicalis
JGI 4.1

Danio rerio
UPDATED! Zv 6 | Vega

Takifugu rubripes
FUGU 4.0

Tetraodon nigroviridis
TETRAODON 7

Pre! *Gasterosteus aculeatus*
BROAD S1

Pre! *Oryzias latipes*
NEW! MEDAKA 1

Ciona intestinalis
JGI2

Ciona savignyi
UPDATED! CSAV 2.0

Drosophila melanogaster
BDGP 4

Anopheles gambiae
AgamP3

Pre! *Aedes aegypti*
AEDES 1

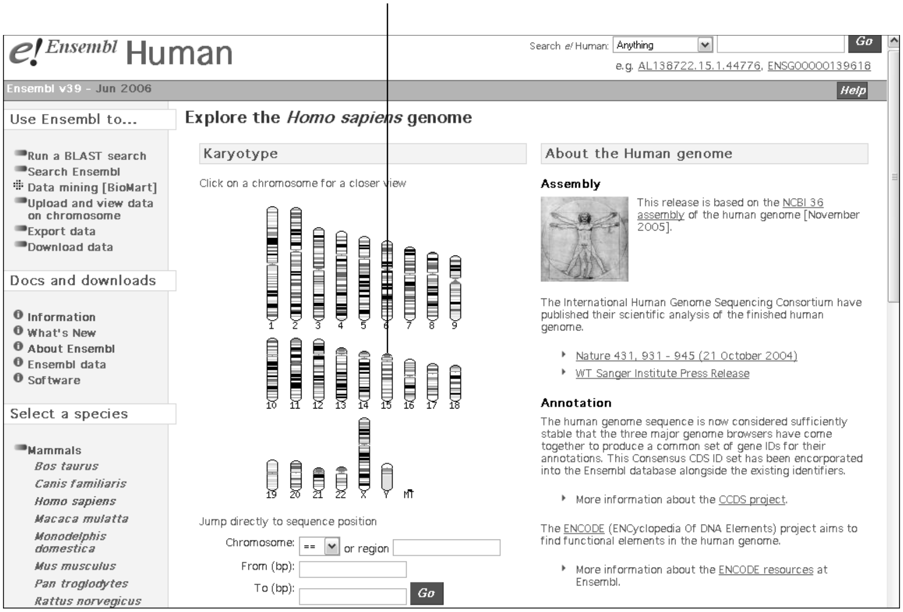
Caenorhabditis elegans
WS 150

Saccharomyces

Figure 3-23: The Ensembl project home page.

2. **Click the Homo sapiens icon.**

   The page that appears (see Figure 3-24) presents a schematic image of the various human chromosomes (numbered according to their size), including the sex chromosomes X and Y, as well as the DNA molecule of the human mitochondria (a small energy-producing organelle present in all human cells). This schematic image is "hot," meaning that clicking a certain area brings up a new window associated with that area.



Click on chromosome 15.

e! Ensembl Human

Search e! Human: Anything Go

e.g. AL138722.15.1.44776, ENSG00000139618

Ensembl v39 - Jun 2006

Help

Use Ensembl to...

- Run a BLAST search
- Search Ensembl
- Data mining [BioMart]
- Upload and view data on chromosome
- Export data
- Download data

Docs and downloads

- Information
- What's New
- About Ensembl
- Ensembl data
- Software

Select a species

- Mammals
  - *Bos taurus*
  - *Canis familiaris*
  - *Homo sapiens*
  - *Macaca mulatta*
  - *Monodelphis domestica*
  - *Mus musculus*
  - *Pan troglodytes*
  - *Rattus norvegicus*

Explore the *Homo sapiens* genome

Karyotype

Click on a chromosome for a closer view

Jump directly to sequence position

Chromosome: == or region

From (bp):

To (bp):

Go

About the Human genome

Assembly

This release is based on the NCBI 36 assembly of the human genome [November 2005].

The International Human Genome Sequencing Consortium have published their scientific analysis of the finished human genome.

- Nature 431, 931 - 945 (21 October 2004)
- WT Sanger Institute Press Release

Annotation

The human genome sequence is now considered sufficiently stable that the three major genome browsers have come together to produce a common set of gene IDs for their annotations. This Consensus CDS ID set has been encorporated into the Ensembl database alongside the existing identifiers.

- More information about the CCDS project.

The ENCODE (ENCyclopedia Of DNA Elements) project aims to find functional elements in the human genome.

- More information about the ENCODE resources at Ensembl.

Figure 3-24: Starting your journey in the human genome.

3. **Click anywhere on the chromosome 15 picture.**

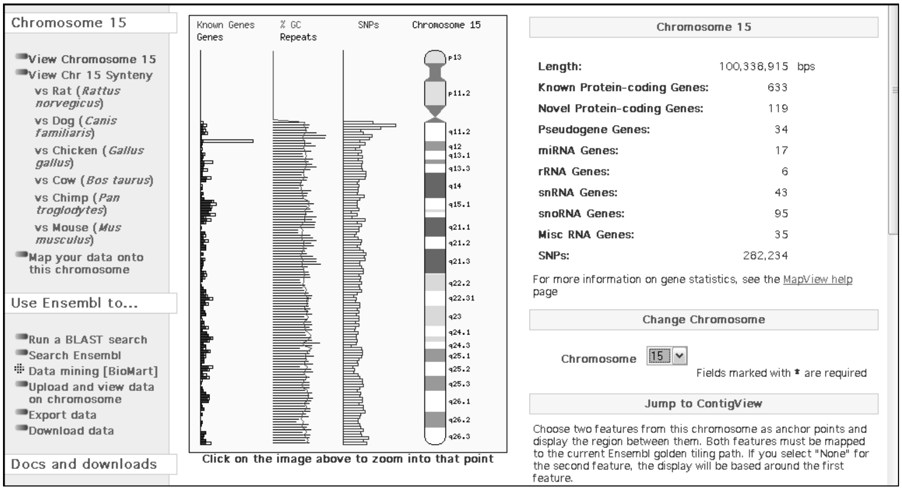
   This lands you in the Chromosome 15 data subset (Figure 3-25), within which we can now ask specific questions, such as: Where is the dUTPase gene? We can do this the hard way by clicking in the region of the q21.1 band and then manually scanning the region. (We see this information earlier in this chapter — remember Figure 3-8?) An easier way is to use the Feature Locator at the bottom right corner of the page, below the Jump to ContigView yellow banner, as shown in Figure 3-26. To do so, follow these steps:

   a. **Select Gene in the From pull-down menu.**

   b. **Enter DUT in the search window, then click the red Go button.**

      You are now looking at a complex output, giving you the structure and the genomic context of the DUT gene at increasing resolution from top to bottom: a schematic overview (1 Mbp range) (see Figure 3-27), a detailed view (10-kbp range) (see Figure 3-28), and even a base-pair view (100-bp range).

      Mousing over the display at various levels lets you examine any chromosome location (down to the level of a gene and its neighbors), study the internal structure and the alternative forms of expression of this specific gene, and assess the eventual consequences of single-nucleotide polymorphisms within that gene (SNP).



Chromosome 15

Length:	100,338,915 bps
Known Protein-coding Genes:	633
Novel Protein-coding Genes:	119
Pseudogene Genes:	34
miRNA Genes:	17
rRNA Genes:	6
snRNA Genes:	43
snoRNA Genes:	95
Misc RNA Genes:	35
SNPs:	282,234

For more information on gene statistics, see the MapView help page

Chromosome 15

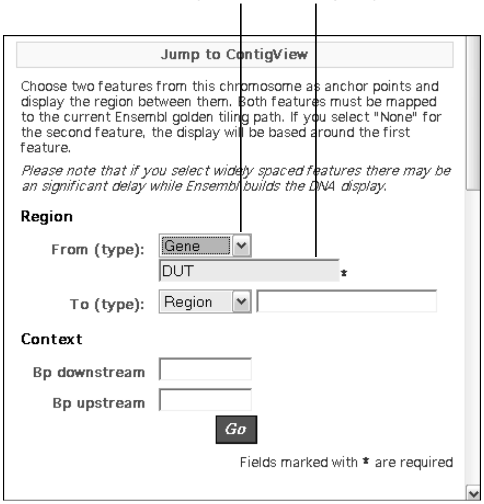
Fields marked with * are required

Choose two features from this chromosome as anchor points and display the region between them. Both features must be mapped to the current Ensembl golden tiling path. If you select "None" for the second feature, the display will be based around the first feature.

Figure 3-25: Focusing on a given genome data subset: Human Chromosome 15.

Select gene.

Input gene name.



Choose two features from this chromosome as anchor points and display the region between them. Both features must be mapped to the current Ensembl golden tiling path. If you select "None" for the second feature, the display will be based around the first feature.

Please note that if you select widely spaced features there may be an significant delay while Ensembl builds the DNA display.

Region

From (type): Gene
DUT *

To (type): Region

Context

Bp downstream

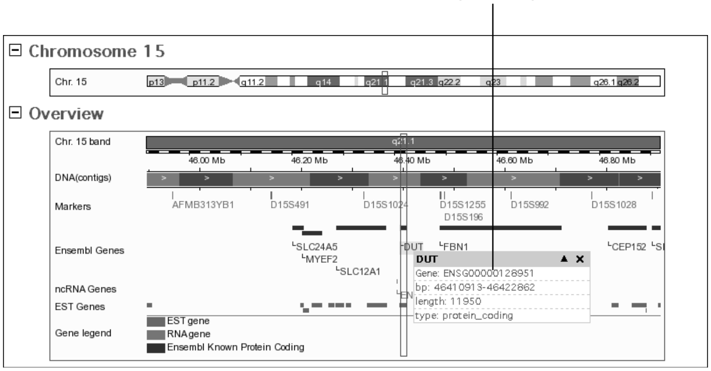
Bp upstream

Fields marked with * are required

Figure 3-26: Looking for the Human dUTPAse gene: DUT.

Getting a complete ID card on the Human DUT gene

Notice that in the overview display (refer to Figure 3-27), the DUT gene name is highlighted. Clicking the gene name here reveals a label with its Ensembl gene number, its location, length, and type (protein coding). In turn, clicking the gene number returns an exhaustive ID card where everything you ever wanted to know about this gene can be found, either explicitly, or as one of the zillions of links to relevant entries in other databases. Check out Figure 3-29 to see how the Human DUT ID card looks like.



Click here to get the gene ID card.

DUT

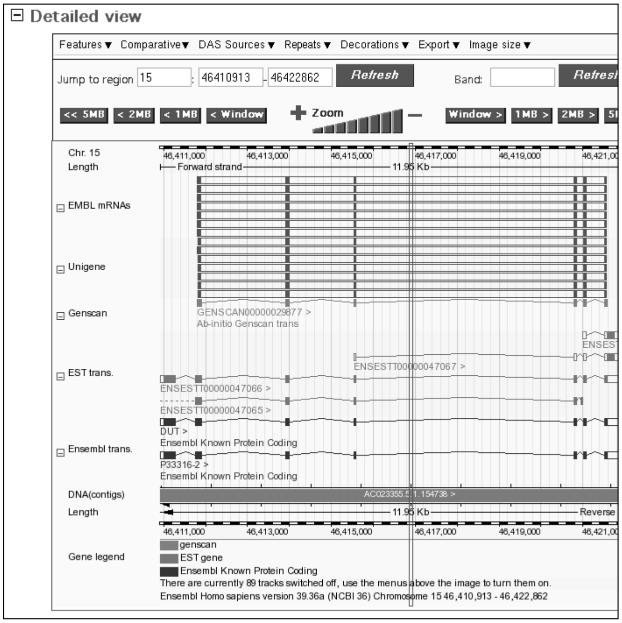
Gene: ENSG00000128951

bp: 46410913-46422862

length: 11950

type: protein_coding

Figure 3-27: The Human DUT gene location: 1-Mbp range overview.



Jump to region 15 : 46410913 - 46422862

There are currently 89 tracks switched off, use the menus above the image to turn them on.

Ensembl Homo sapiens version 39.36a (NCBI 36) Chromosome 15 46,410,913 - 46,422,862

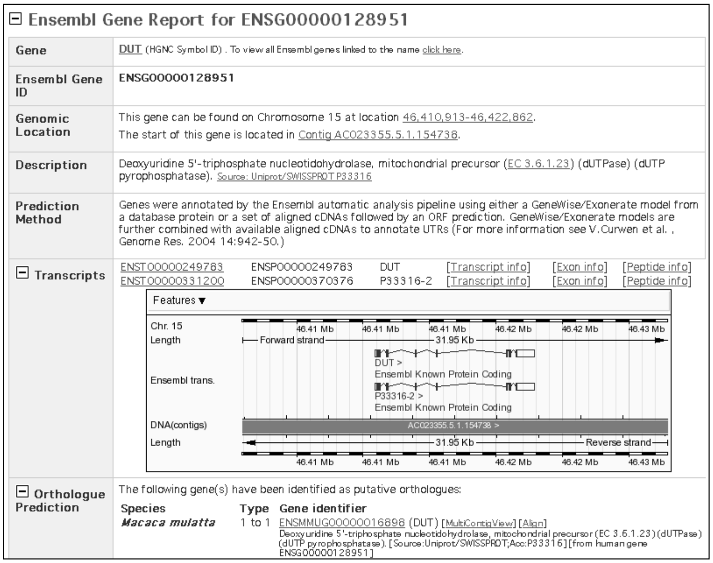
Figure 3-28: The Human DUT gene location: 10-kb range detailed view.

Finding disease genes with coding SNPs using BioMart

There's more to Ensembl fun than just random browsing. If you're a serious scientist, it can answer meaningful questions very easily, very precisely, and very quickly, even if the output isn't always as pretty as what we showed you in the preceding section. The data-mining system that the Ensembl people designed for this purpose is called BioMart. (Refer to Figure 3-23, upper-left corner.)

Imagine that, as a competent human geneticist, you ask yourself the following questions:

> *Which genes on Chromosome 15 have been linked to a disease?*
>
> *Which SNPs (Single-Nucleotide Polymorphisms) have been identified?*



⊟ Ensembl Gene Report for ENSG00000128951

Gene	DUT (HGNC Symbol ID) . To view all Ensembl genes linked to the name click here.
Ensembl Gene ID	ENSG00000128951
Genomic Location	This gene can be found on Chromosome 15 at location 46,410,913-46,422,862. The start of this gene is located in Contig AC023355.5.1.154738.
Description	Deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial precursor (EC 3.6.1.23) (dUTPase) (dUTP pyrophosphatase). Source: Uniprot/SWISSPROT P33316
Prediction Method	Genes were annotated by the Ensembl automatic analysis pipeline using either a GeneWise/Exonerate model from a database protein or a set of aligned cDNAs followed by an ORF prediction. GeneWise/Exonerate models are further combined with available aligned cDNAs to annotate UTRs (For more information see V.Curwen et al. , Genome Res. 2004 14:942-50.)
⊟ Transcripts	ENST00000249783 ENSP00000249783 DUT [Transcript info] [Exon info] [Peptide info] ENST00000331200 ENSP00000370376 P33316-2 [Transcript info] [Exon info] [Peptide info]
⊟ Orthologue Prediction	The following gene(s) have been identified as putative orthologues: Species: *Macaca mulatta* — Type: 1 to 1 — Gene identifier: ENSMMUG00000016898 (DUT) [MultiContigView] [Align] Deoxyuridine 5'-triphosphate nucleotidohydrolase, mitochondrial precursor (EC 3.6.1.23)(dUTPase) (dUTP pyrophosphatase). [Source:Uniprot/SWISSPROT;Acc:P33316][from human gene ENSG00000128951]

Figure 3-29: Top part of the Ensembl ID card for the Human DUT gene.

If you're just dying to find a candidate gene for a disease you've just genetically mapped on chromosome 15, the first question makes a lot of sense.

To make things even more interesting, we can look for *coding SNPs* — the impetus for our second question. (SNPS involve nucleotide changes that may alter the sequence and possibly the shape and/or function of the corresponding protein product.)

Now, both of these questions are real doozies. Believe it or not, you can answer them in less than a minute with Ensembl! Follow these steps to see how:

1. **Point your browser to** `www.ensembl.org/`**.**

   The Project Ensembl home page appears.

2. **Click the `Data Mining` link in the upper-left corner of the page. (Refer to Figure 3-23.)**

   Your browser displays the starting page MartView page, which is already set up for working with the last version of the Human genome data (otherwise select the dataset).

3. **Simply proceed by clicking the Next button.**

   The new page looks formidable, with a zillion selectable search options. Don't panic; we'll get you through it in no time!

4. **At the top of the form (Region set up), check the Chromosome box, and select 15 from the pull-down menu.**

5. **In the following section (Gene set up), check the top box, and select With Disease Association from the accompanying pull-down menu. Then check the Gene Type box, and select protein_coding from the associated pull-down menu.**

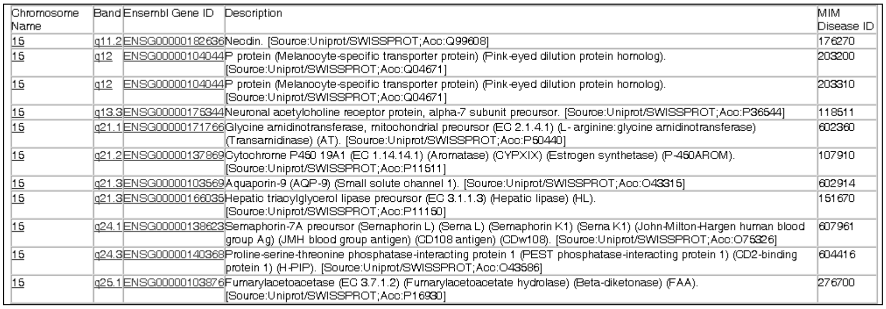
6. **Scroll down the form until you reach the SNP set up area; check the Coding box and then click the Next button at the bottom of the page.**

7. **In the Output form that appears, check the Chromosome Name and Band boxes, and then check the Ensembl Gene ID, Description and MIM Disease ID boxes in the next section.**

   MIM stands for *M*endelian *I*nheritance in *M*an, and is the leading human-genetic-disease database.

10. **Proceed to the bottom of the form and click the Export button.**

    In no time at all, Ensembl returns a table that looks like Figure 3-30. Our BioMart search identified many different genes known to be associated with various diseases, and for which sequence polymorphisms inducing protein changes have been documented. These will be our prime candidates for further research. As before (refer to Figure 3-29), clicking the Ensembl Gene ID will produce a comprehensive information card for these genes, allowing you to quickly figure out whether their functions are related to the disease symptoms.



Chromosome Name	Band	Ensembl Gene ID	Description	MIM Disease ID
15	q11.2	ENSG00000182636	Necdin. [Source:Uniprot/SWISSPROT;Acc:Q99608]	176270
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT;Acc:Q04671]	203200
15	q12	ENSG00000104044	P protein (Melanocyte-specific transporter protein) (Pink-eyed dilution protein homolog). [Source:Uniprot/SWISSPROT;Acc:Q04671]	203310
15	q13.3	ENSG00000175344	Neuronal acetylcholine receptor protein, alpha-7 subunit precursor. [Source:Uniprot/SWISSPROT;Acc:P36544]	118511
15	q21.1	ENSG00000171766	Glycine amidinotransferase, mitochondrial precursor (EC 2.1.4.1) (L- arginine:glycine amidinotransferase) (Transamidinase) (AT). [Source:Uniprot/SWISSPROT;Acc:P50440]	602360
15	q21.2	ENSG00000137869	Cytochrome P450 19A1 (EC 1.14.14.1) (Aromatase) (CYPXIX) (Estrogen synthetase) (P-450AROM). [Source:Uniprot/SWISSPROT;Acc:P11511]	107910
15	q21.3	ENSG00000103569	Aquaporin-9 (AQP-9) (Small solute channel 1). [Source:Uniprot/SWISSPROT;Acc:O43315]	602914
15	q21.3	ENSG00000166035	Hepatic triacylglycerol lipase precursor (EC 3.1.1.3) (Hepatic lipase) (HL). [Source:Uniprot/SWISSPROT;Acc:P11150]	151670
15	q24.1	ENSG00000138623	Semaphorin-7A precursor (Semaphorin L) (Sema L) (Semaphorin K1) (Sema K1) (John-Milton-Hargen human blood group Ag) (JMH blood group antigen) (CD108 antigen) (CDw108). [Source:Uniprot/SWISSPROT;Acc:O75326]	607961
15	q24.3	ENSG00000140368	Proline-serine-threonine phosphatase-interacting protein 1 (PEST phosphatase-interacting protein 1) (CD2-binding protein 1) (H-PIP). [Source:Uniprot/SWISSPROT;Acc:O43586]	604416
15	q25.1	ENSG00000103876	Fumarylacetoacetase (EC 3.7.1.2) (Fumarylacetoacetate hydrolase) (Beta-diketonase) (FAA). [Source:Uniprot/SWISSPROT;Acc:P16930]	276700

Figure 3-30: Output of your first BioMart query.

The range and complexity of the questions you can address through the Ensembl BioMart resource is truly impressive. We really encourage you to spend some time playing with it, even though it isn't as visual as the zooming tools we showed you in previous sections.

Chapter 4

Using Protein and Specialized Sequence Databases

In This Chapter

- Exploring protein maturation
- Understanding a UniProtKB/Swiss-Prot entry from the top down
- Finding out about detailed protein biochemistry
- Discovering the function of your protein
- Hunting up useful protein-related information resources

> *Mankind is a catalyzing enzyme for the transition from a carbon-based to a silicon-based intelligence.*
>
> — Gerard Bricogne

We've focused this chapter to finding information on protein sequences. The databases on proteins don't limit themselves to providing sequences. In this chapter, we show you that by mastering the — sometimes tangled — network of Web links available to you on the Internet, you can relate proteins to their gene sequences, to their functions, and even to their 3-D structures.

Don't let that abundance and diversity of information fool you. As it is for the nucleotide sequence world, where everything revolves around one central resource (that is, GenBank), most of the links between genes, proteins, and functions rely on UniProtKB/Swiss-Prot, the central resource on annotated protein sequences co-funded by the Swiss Institute of Bioinformatics (SIB) and the European Bioinformatics Institute (EBI). In this chapter, we show you how to extract every bit of information you need from this database.

Swiss-Prot: A personal vision of the protein world

Although GenBank and Swiss-Prot occupy similarly central roles for the international community of biologists, their overall philosophies aren't quite the same. GenBank, as a primary sequence repository, obeys a relatively strict historical point of view. In GenBank, the authors have full authority over the content of the entries they submit. GenBank annotators are only responsible for the recently introduced RefSeq entries — the ones they derive from their own expert analysis of the community-submitted entries. However, the RefSeq entries never replace the original GenBank entries, and all these layers of information are maintained side by side.

In contrast, Swiss-Prot is not a repository database but a derived information resource, conveying the vision of its head and founder, Amos Bairoch (helped out by a group of experts). As a consequence, the only truly current Swiss-Prot version is the one on Amos' portable computer. This idea of "personal vision" also means that Amos Bairoch doesn't need anybody's permission to correct or change a Swiss-Prot entry on the spot. To do this, Amos simply needs to be convinced by an expert, or by his own evaluation of the literature, that a change is necessary. Believe us, we have seen this happening on our own kitchen table!

The benefit of such a philosophy is obviously great when it comes to flexibility; blatant errors can be removed very quickly while hot discoveries are incorporated immediately. This is the reason why Swiss-Prot is considered the best-annotated protein database, and why it occupies such a pivotal role in molecular biology and genomics. The downside involves the (inevitable) upper limit of what the best researcher in the world can do (or supervise), even when helped by a large team of annotators — at the European Bioinformatics Institute (EBI) in Hinxton as well as the Swiss Institute of Bioinformatics (SIB) in Geneva — and a circle of experts. These days, Swiss-Prot has troubles coping with the present rate of new (nucleotide) sequence determination and is falling behind in terms of completeness.

To alleviate this problem, a Swiss-Prot buffer has been created that is called TrEMBL (for automatic *TR*anslation of *E*uropean *M*olecular *B*iology *L*aboratory nucleotide sequences). TrEMBL entries are generated at the EBI from GenBank submissions and annotated mostly automatically, using sequence similarity as a main criterion. Upon visual inspection, manual correction, and final approval by Amos Bairoch, TrEMBL entries are then converted into *bona fide* Swiss-Prot entries.

The idea that such a key worldwide information resource rests on a single's man shoulders may come as a shock to you. However, this sort of thing has been quite common since the early days of bioinformatics. Among other famous examples of one-man shows, we can cite Elvin Kabat 's Immunoglobulin sequence database, Richard J. Roberts' Restriction Enzyme Database, or Victor McKusick's database of human genetic diseases (Online Mendelian Inheritance in Man, OMIM). To paraphrase Sir Winston Churchill: "In the field of bioinformatics, rarely have so many owed so much to so few!"

You have many ways of landing on the Swiss-Prot database (or the UniProt knowledge base, as they also call it now!). Most genomic databases, genome browsers, or sites running a sequence retrieval system (SRS) link you directly to the relevant Swiss-Prot entry. If you want to know how to query Swiss-Prot to find an entry by keyword, go to Chapter 2, where we explain how to do this in some detail.

The focus of this chapter is on giving you a good understanding of the content of a Swiss-Prot entry so that you know exactly what is what in the very exhaustive annotation of this database. We then show you how you can use this annotation as a starting point to visit more specialized databases. You can use these databases to further your investigations of the function or structure of your protein.

We start this chapter by giving you some general ideas on how the cell turns a native protein sequence into a mature protein. These post-translational modifications are very important for the protein function. Finding out about them is one of the main reasons you want to use a protein database. (See Chapter 6 if you want to find out how to predict these modifications.)

From Translated ORFs to Mature Proteins

Because sequencing DNA molecules is so much easier and cheaper than sequencing proteins, most of the amino-acid sequences we know do in fact correspond to computer translations of open reading frames (ORFs) detected through the analysis of genomic data. Most of the corresponding proteins have never actually been isolated by anybody. This is why protein specialists hate us molecular biologists and bioinformaticists when we keep talking about translated ORFs as if they were genuine proteins. For a while, at the beginning of the genomic era, the easy thing to do was to consider these protein specialists a bunch of grumpy old men and simply ignore them.

However, after genomes were sequenced, people's interest in proteins came back in a large-scale format known as proteomics. *Proteomics* is the scientific field dealing with the visualization and quantification of the set of protein molecules present in a given tissue or organism. Proteomics analysis brought a rapid accumulation of data — and quickly demonstrated that protein specialists had very good reasons to be angry with us.

ORFs: What you see is NOT what you get

When biologists perform a 2-D gel electrophoresis (a tricky experiment separating protein molecules according to their mass in one direction and their charge in the other), real proteins almost never behave as you'd expect given the computer translation. In the world of ORFs and proteins, *what you see is NOT what you get.*

The reason is that, when translated from RNA, the nascent amino-acid chain can be heavily modified on its way to becoming a mature protein. Even simple physico-chemical properties of a mature protein — such as size, molecular weight, or isoelectric point — are hard to predict if you know only the computer-generated amino-acid sequence. This complex process of protein maturation is referred to as *post-translational modification.* Figure 4-1 summarizes it.

Protein maturation may include any combination of the following stages:

- Cuts within the amino-acid chain
- Removal of fragments of the amino-acid chain (this is the case for insulin)
- Chemical modifications of specific amino acids (methylation, for example)
- Addition of lipid molecules (myristoylation, for example)
- Addition of glycosidic (sugar) molecules (glycosylation, for example)

A major role for a *protein* database (in contrast with an ORF sequence collection) is to display this type of information, when it is available from experimental data or is predicted using various computational techniques. (For more on such techniques, see Chapter 6). Hence post-translational modifications make up a sizeable portion of the protein features recorded in Swiss-Prot entries.

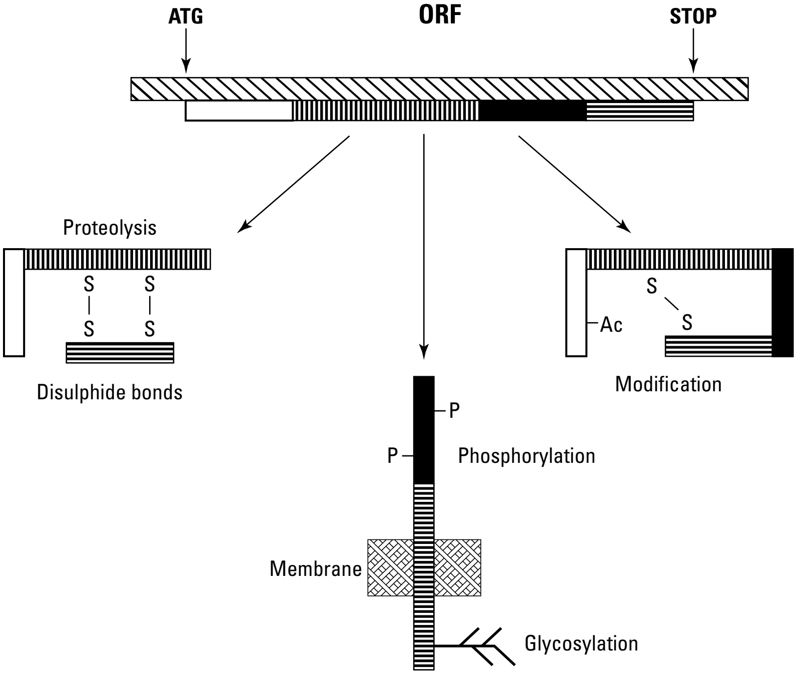


Figure 4-1: Possible modifications encountered during protein maturation.

A personal final destination for each protein

In order to correctly accomplish their functions, proteins have to reach the right destination in the organism or within the cell. As it is translated, the peptide chain may expose a variety of highly specific sequence signals — "ZIP codes," if you will — which the cell then uses to direct the protein to the appropriate compartment (in or out of the cell). This sorting always involves the transport of the protein across one or several membranes and is also referred to as *translocation.* The final activities and destinations of a protein include

- Getting attached to the cell membrane
- Being secreted outside the cell
- Being transported into the periplasm (for bacteria)
- Being transported to the mitochondria or any other organelle
- Being transported into the cell nucleus

Because knowing the final compartment where a protein ends up is important in understanding its function, this information (proven or predicted) is one of the important features recorded in protein databases like Swiss-Prot.

A combinatorial diversity of folds and functions

The most important step in turning a newly synthesized peptide chain into a functional protein is the folding of this chain into a compact and stable 3-D structure. Except for small proteins (fewer than 100 amino acids), the final protein structure generally consists of several relatively independent domains. You can imagine these domains as a basic set of fairly rigid LEGO bricks. Nature can assemble these bricks to produce the immense variety of existing proteins. For instance, given a basic set of three domains (A, B, C), you can end up with Protein AAA or AB or BCC or BAC, and so on.

Most natural proteins are made of combinations of one to ten domains picked from a set of a few thousands. Despite significant sequence variations, the domains are identifiable by their *scaffold sequence signatures* — the motifs in the protein amino-acid texts that remain recognizable despite a zillion years of divergent evolution. The recognition and the definition of protein domains is a major research topic of bioinformatics. (See Chapters 6, 7, 9 and 11.) The domain architecture underlying a particular protein sequence is important to know because it gives hints about the protein's possible 3-D structure — and suggests its potential biochemical or cellular function.

For this reason, a significant part of each Swiss-Prot entry is devoted to the description of the domain organization of proteins, as identified or predicted by all available methods.

Reading a Swiss-Prot Entry

Despite the complicated concepts that their descriptions rely on, proteins are much simpler objects than genes:

- Proteins correspond to relatively small sequences (350 amino acids long, on the average).
- Unlike genes, proteins have clear beginnings and clear ends.
- Proteins are defined on a single strand.
- Whatever modifications occur between the ORF sequence and the mature protein, the amino acids they contain remain in the same order.

Given all these reasons, you'd probably expect Swiss-Prot entries to be easier to read and understand than GenBank's — and you'd be right!

Like a GenBank entry — which we cover with great flair in Chapter 3 — a Swiss-Prot entry is divided into several main sections:

- The general information part
- The bibliographic information part
- The functional information part
- The feature table
- The sequence part

Using the human *epidermal growth factor receptor* (EGFR) as an example, the following sections show you how to decipher a Swiss-Prot entry from top to bottom, step by step (or line by line).

Deciphering the EGFR Swiss-Prot entry

The EGF receptor is a rather complex molecule. We use it here because it illustrates remarkably well the wealth of information you can glean from a Swiss-Prot entry.

Chapter 2 shows you more sophisticated ways of querying Swiss-Prot, but for our purposes here, we can use the following basic search:

1. **Point your browser to** `www.expasy.ch/sprot/`**.**

   The UniProtKB/Swiss-Prot page of the ExPASy server appears (refer to Figure 2-10, Chapter 2, to remember what it looks like).

2. **Type the Swiss-Prot ID** P00533 **in the Search window (top of the page).**

   This is the entry corresponding to the human receptor of the epidermal growth factor. We chose it because this protein is hot! It is an oncogene (a trigger for cancer), and it exhibits a fairly detailed annotation. Because more people work on hot proteins, they're usually better annotated than the more mundane ones.

3. **Click the Go button.**

   The UniProtKB/Swiss-Prot P00533 entry page appears. If you want to print this page, click the Printer-Friendly View button at the top right.

REMEMBER

Because of the large amount of information available that deals with this particular protein, this entry is probably among the most complicated you'll ever come across when interacting with Swiss-Prot. Most other entries don't exhibit a third of the keywords and fields you can find here.

General information about the entry

You're now ready to go through this fairly complex entry step by step. Blue banners separate the main sections across the screen. The first one is titled Entry Information and contains the following six fields:

- **Entry Name: *EGFR_HUMAN*.** This Swiss-Prot identifier gives a quick idea of what the entry is all about. Here, the name is quite explicit, but don't count on that — such clarity isn't always the case! For instance, the fly lysozyme entry name is *LYSA_DROME*, certainly more ambiguous and more difficult to remember (DROME stands for *DRO*sophila *ME*lanogaster).

  WARNING!

  As a general rule, the entry name is not supposed to be meaningful, nor stable. As a consequence, you're strongly advised *not* to use it for cross-reference purposes, or as a sequence identifier in an article. Instead, you should use the primary accession number that appears on the next line.

- **Primary Accession Number: *P00533*:** This is the truly unique and stable identifier of this sequence. You must use this number when referencing the entry in your work.

  Accession numbers are great because they allow you to retrace the database history of a given protein. For instance, you can imagine that the EGF-receptor protein, being an unusually long protein, was first described as a partial sequence. Then people worked further to determine longer fragments of this protein, and so on. At each step, they published a paper and submitted a new version of the EGF-receptor protein (perhaps in the process correcting earlier mistakes). Unlike GenBank, Swiss-Prot keeps only one version of the EGF receptor and summarizes our knowledge up to this day.

  Ancient accession numbers never die; they just get stored in the secondary accession number field.

- **Secondary Accession Numbers: O00688, *P06268, Q14225, et al*:** Secondary accession numbers make sure that when you read an old paper mentioning Q12225 or P06268 as accession numbers for EGFR, you can still use the old reference to find the most recent one. See for yourself and try to use one of these older numbers in the Search window.

  Whatever secondary accession number you type in, you always fall back on the current one: P00533. This is a clean and simple way to remove old entries while keeping them unambiguously linked to the state-of-the-art version.

  The next lines in the General information are self-explanatory.

- **Integrated into Swiss-Prot on: *July 21, 1986.***
- **Sequence was last modified on: *November 1, 1997.***

  This suggests that it took ten years to get the whole sequence right!

- **Annotations were last modified on: *July 25, 2006.***

  Clearly, a lot of work is still going on about this important protein.

TIP

Notice that, for the whole entry, the field definition (the left part of each line) is an active link that directs you to the relevant part of the Swiss-Prot manual. Reading a few of those might earn you some respect from the local bioinformatics gurus!

Name and origin of the protein

At this point, we encounter a new section (introduced by the blue banner): Name and Origin of the Protein. It contains the following five fields:

- **Protein name: *Epidermal growth factor receptor [Precursor]:*** This field contains general descriptive information about the sequence. It is generally sufficient to identify the protein precisely. If the sequence is a fragment, this field indicates it.

  Here, the `[Precursor]` attribute indicates that the sequence must be cleaved to form the mature form of the protein.

- **Synonyms: *Receptor tyrosine-protein kinase ErbB-1*, *EC 2.7.10.1:*** The E.C. number (2.7.10.1) encodes the biochemical reaction (tyrosine-protein kinase) that this protein performs. E.C. stands for Enzyme Nomenclature Committee, and we give you a bit more information about it in Table 4-1.

  This little line doesn't look very impressive, but it can be the starting point of a fascinating journey toward a complete understanding of this protein enzymatic function. For example, click the 2.7.10.1 link to see the NiceZyme view of the enzyme, as shown in Figure 4-2. The links on this page lead to various types of information; you can

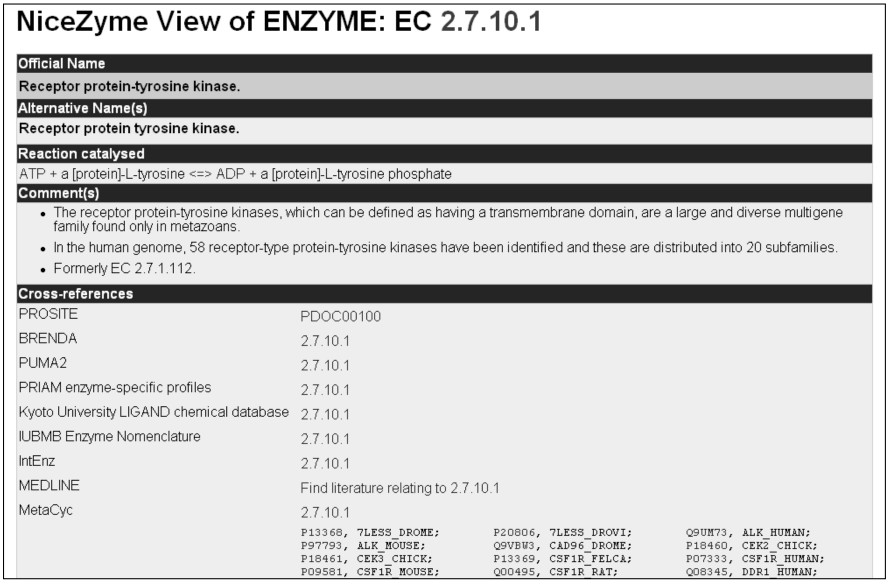
  - Locate your protein within a chart of all cellular metabolic pathways
  - Get a description and sequence profile of the family it belongs to
  - Get the chemical formulas of the biochemical compounds it can interact with

- **Gene name: *EGFR:*** This field is useful when you need to interact with nucleotide sequence and genome databases. It can help keep queries clear and unambiguous. (For more on gene names, see Chapter 3.)

- **From: *Homo sapiens (Human) [TaxID:9606]:*** The `From` field defines precisely the origin of the protein. The species designation consists, in most cases, of the Latin genus and species designation followed by the English name (in parentheses). For viruses, only the common English name is given. Clicking the Taxonomic ID 9606 link gets you to the *H.sapiens* page of the NEWT taxonomy database maintained by the UniProtKB group (`www.ebi.ac.uk/newt/index.html`).

- **Taxonomy: *Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; ...,* etc:** This field lists the taxonomic classification of the source organism. The taxonomic classification used is that maintained at the National Center for Biotechnology Information (NCBI). Each group name is a link to a list of all the Swiss-Prot entries from that group.

  This field is handy if you want to quickly gather all the protein sequences of a given species for your research projects.



NiceZyme View of ENZYME: EC 2.7.10.1

Official Name

Receptor protein-tyrosine kinase.

Alternative Name(s)

Receptor protein tyrosine kinase.

Reaction catalysed

ATP + a [protein]-L-tyrosine <=> ADP + a [protein]-L-tyrosine phosphate

Comment(s)

- The receptor protein-tyrosine kinases, which can be defined as having a transmembrane domain, are a large and diverse multigene family found only in metazoans.
- In the human genome, 58 receptor-type protein-tyrosine kinases have been identified and these are distributed into 20 subfamilies.
- Formerly EC 2.7.1.112.

Cross-references

PROSITE	PDOC00100
BRENDA	2.7.10.1
PUMA2	2.7.10.1
PRIAM enzyme-specific profiles	2.7.10.1
Kyoto University LIGAND chemical database	2.7.10.1
IUBMB Enzyme Nomenclature	2.7.10.1
IntEnz	2.7.10.1
MEDLINE	Find literature relating to 2.7.10.1
MetaCyc	2.7.10.1

```
P13368, 7LESS_DROME;      P20806, 7LESS_DROVI;      Q9UM73, ALK_HUMAN;
P97793, ALK_MOUSE;        Q9VBW3, CAD96_DROME;      P18460, CEK2_CHICK;
P18461, CEK3_CHICK;       P13369, CSF1R_FELCA;      P07333, CSF1R_HUMAN;
P09581, CSF1R_MOUSE;      Q00495, CSF1R_RAT;        Q08345, DDR1_HUMAN;
```

Figure 4-2: NiceZyme view of Swiss-Prot entry P00533.

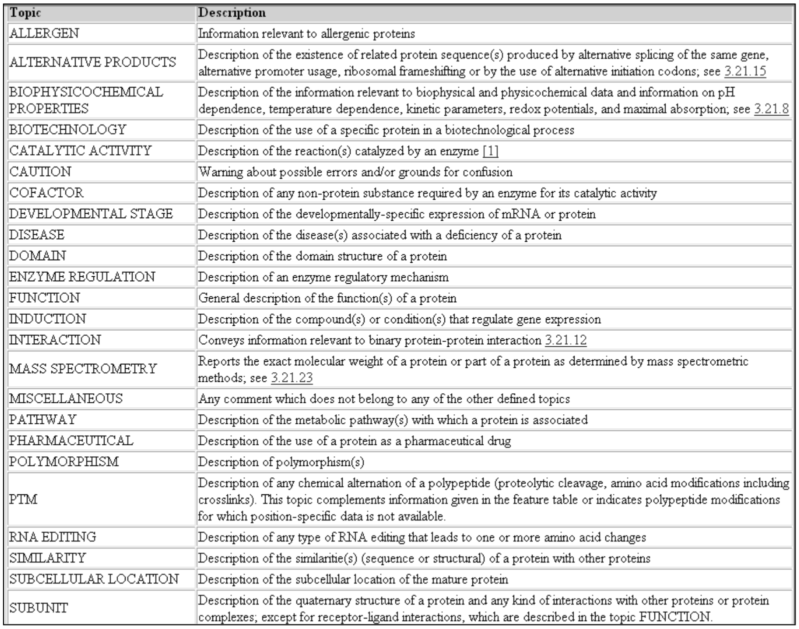
The References

The next section of the P00533 entry is the References section — and it really is as simple as it looks: a list of the bibliographical references used to build the current entry. In this example, it includes the various studies that have contributed to the sequencing, to the definition of several isoforms of the protein as expressed in different tissues or tumors, and to the mapping of various post-translational modifications, disulphide bonds, or ligand-binding sites. This protein's importance can be measured by the number of references here — 34 is pretty impressive.

For each of the published citations, Swiss-Prot provides the relevant NCBI/PubMed identifier and for most of them a direct link to the article abstract.

The Comments

The Comments section holds any useful information that just doesn't fit anywhere else. The comments are free texts grouped together in comment blocks; a block contains one or more comment lines and is introduced by a topic keyword. The topic keywords themselves are pretty wide ranging, as the listing in Figure 4-3 makes clear. (You can get to this listing — part of the Swiss-Prot manual — by clicking the Comments heading.)



Topic	Description
ALLERGEN	Information relevant to allergenic proteins
ALTERNATIVE PRODUCTS	Description of the existence of related protein sequence(s) produced by alternative splicing of the same gene, alternative promoter usage, ribosomal frameshifting or by the use of alternative initiation codons; see 3.21.15
BIOPHYSICOCHEMICAL PROPERTIES	Description of the information relevant to biophysical and physicochemical data and information on pH dependence, temperature dependence, kinetic parameters, redox potentials, and maximal absorption; see 3.21.8
BIOTECHNOLOGY	Description of the use of a specific protein in a biotechnological process
CATALYTIC ACTIVITY	Description of the reaction(s) catalyzed by an enzyme [1]
CAUTION	Warning about possible errors and/or grounds for confusion
COFACTOR	Description of any non-protein substance required by an enzyme for its catalytic activity
DEVELOPMENTAL STAGE	Description of the developmentally-specific expression of mRNA or protein
DISEASE	Description of the disease(s) associated with a deficiency of a protein
DOMAIN	Description of the domain structure of a protein
ENZYME REGULATION	Description of an enzyme regulatory mechanism
FUNCTION	General description of the function(s) of a protein
INDUCTION	Description of the compound(s) or condition(s) that regulate gene expression
INTERACTION	Conveys information relevant to binary protein-protein interaction 3.21.12
MASS SPECTROMETRY	Reports the exact molecular weight of a protein or part of a protein as determined by mass spectrometric methods; see 3.21.23
MISCELLANEOUS	Any comment which does not belong to any of the other defined topics
PATHWAY	Description of the metabolic pathway(s) with which a protein is associated
PHARMACEUTICAL	Description of the use of a protein as a pharmaceutical drug
POLYMORPHISM	Description of polymorphism(s)
PTM	Description of any chemical alternation of a polypeptide (proteolytic cleavage, amino acid modifications including crosslinks). This topic complements information given in the feature table or indicates polypeptide modifications for which position-specific data is not available.
RNA EDITING	Description of any type of RNA editing that leads to one or more amino acid changes
SIMILARITY	Description of the similaritie(s) (sequence or structural) of a protein with other proteins
SUBCELLULAR LOCATION	Description of the subcellular location of the mature protein
SUBUNIT	Description of the quaternary structure of a protein and any kind of interactions with other proteins or protein complexes; except for receptor-ligand interactions, which are described in the topic FUNCTION.

Figure 4-3: Key topics in the Comments section of Swiss-Prot.

The Comments section of entry P00533 is particularly rich, as it includes information about the following topics:

- **FUNCTION:** Receptor for EGF, involved in the control of cell growth
- **CATALYTIC ACTIVITY:** Tyrosine-protein kinase
- **SUBUNIT:** Other proteins it forms stable complexes with
- **INTERACTION:** Other proteins it interacts with
- **SUBCELLULAR LOCATION:** Cellular-membrane protein
- **ALTERNATIVE PRODUCTS:** The isoforms from splice variants
- **TISSUE SPECIFICITY:** Placenta, ovarian cancer
- **PTM:** List some of the Post Translational Modifications
- **DISEASE:** Defects in EGFR are associated with lung cancer
- **MISCELLANEOUS:** Used here to describe the molecular mechanism (dimerization) underlying the function of the protein
- **SIMILARITY:** With other sequences of the EGF receptor subfamily
- **WEB RESOURCE:** A link to GeneTests, an NIH-funded database on clinically available genetic tests

The content of the Comments section is often sufficient to decide whether a match (with BLAST or otherwise) of your query with a particular Swiss-Prot entry makes sense or not.

Let's skip the Copyright notice — the stuff for the lawyers out there — and jump into the big Cross-References section that comes next.

The Cross-References

The Cross-References section contains links to entries in other databases that contain some information about our protein.

TIP

Most of the links in this section take you to new databases. With all this Web jumping about from database to database, it can be easy to get lost and lose your original Swiss-Prot entry. To avoid this confusion, when you browse the Cross-References section, open the links in a new window: Click the link with the *right* button of your mouse (not the left, as you usually do) and choose Open in a New Window from the context menu that appears.

There are bunches of fields in the Cross-References section. The following list will help you keep at least the major ones straight:

- **EMBL:** This field contains all necessary links with the nucleotide sequences world. As we point out in Chapter 3, numerous GenBank, EMBL, and DDBJ entries can be related to a single protein sequence.

  TIP

  Click the CoDingSequence link to send a query to the EBI SRS server for finding the CDS (*CoD*ing *S*egment, the part of the nucleotide sequence that precisely encodes the protein from start to stop) of these entries.
- PIR: This historical field contains the accession numbers of the corresponding entries in the late Protein Information Resource (PIR), now discontinued, but incorporated in the UniProtKB consortium.
- UniGene: A link to NCBI's gene expression database (see also the CleanEx field for linking to DNA chip experimental data).
- **PDB:** This field contains a link to a sequence homologue of the current query, for which 3-D structural information (X-ray crystal structures) is available. Here, this is Swiss-Prot entry P11362, the basic fibroblast growth factor 1, for which several crystal structures exist in the protein Databank (PDB). Click the PDB link to see the relevant page. (The PDB ID here is 1FGK.)
- **ModBase:** This field links you to ModBase, a database of theoretically calculated models, not experimentally determined structures. The models may contain significant errors, as pointed out by the authors themselves.

In Chapter 11, we explain all the great things you can do when homologous structural information is available for your protein.

There is no PDB field when no direct or indirect structural information exists. This is the case for a sizeable portion of the entries in Swiss-Prot.

- **DIP and IntAct:** Links to protein-protein interaction databases. DIP stands for *D*atabase of *I*nteracting *P*roteins, a database maintained by UCLA. This is an interesting resource in that it lists all the proteins that have been experimentally shown to interact with our EGF-receptor protein. IntAct (a project from EBI) is built from the information found in the literature.
- **GlycoSuiteDb:** This field is only there when your protein has a known glycosylation pattern. It provides a link to the relevant chemical composition of the sugar moiety. This is a nice site, but the free service is provided only to academics and requires a personal registration, a significant drawback to other researchers.
- **SWISS-2DPAGE:** This field contains a link to the proteomics world. Clicking SWISS-2DPAGE brings you to a collection of 2-D gel electrophoreses for different human tissues or tumors. If your protein has been identified on a gel, the spots are highlighted. This is the case for P00533. Otherwise, all that appears on-screen is a box of its expected location (drawn after its theoretical isoelectric point pI and predicted molecular weight). This is great if you're ready for real experimental work!
- **HGNC:** This field (for human proteins only) provides a link to the official human gene names provided by the HUGO (*HU*man *G*enome *O*rganization) Gene Nomenclature Committee.
- **GeneCards:** This field provides a link to the relevant entry in the GeneCards database maintained by the Weizmann Institute of Science, based in Israel.

  GeneLynx: This field provides a link to the relevant entry of GeneLynx, a collection of hyperlinks to the publicly available resources associated with each human gene.
- **MIM:** This field (mostly for human/mouse proteins) provides a link to the relevant entry in OMIM, the human genetic disease database Online Mendelian Inheritance in Man.
- **Ontologies:** This subsection attempts to list, in a hierarchical manner (called an *ontology*) all the functions the EGFR gene product is known to be associated with. Clicking on the QuickGo view link (at the send of this subsection) gets you a nice summary of all that information.

- **Profiles/domains:** The Cross-References section contains a series of fields that deal with the classification of the protein in families of related sequences based on their sharing of basic domains.

  Many different concurrent classification systems exist. They all use different methods (algorithms) for the definition and detection of the recurrent structural/sequence domains, as well as for the clustering of the proteins in families. The principle behind these classifications and the use you can make of them are well described in Chapters 7, 9, and 11. The relevant keywords are

  - InterPro (summarizing most of the following ones)
  - Pfam
  - Prodom
  - PRINTS
  - SMART
  - PROSITE
  - BLOCKs

  If you have time for only one click, use the InterPro link; it is a summary of all the others (except for BLOCKs). Clicking the Graphical view of domain structure link in the InterPro row shows you at once that many different recurrent domains have been identified in the EGF receptor protein P00533.

- **Ensembl:** This field is only here for human and mouse proteins. It provides a link to the relevant gene report in the Ensembl database, a major human genome resource. (For more on Ensembl, see Chapter 3.)

  TIP

  With the NCBI/PubMed and the PDB links, we believe this is one of the most useful hyperlinks provided in the Swiss-Prot entries for mammalian genes.

- **RZPD-ProtExp and SOURCE:** These useful fields (for the experimentalists) let you identify sources of publicly available starting material (like cDNA clones) in case you decide to work on the *EGFR* gene, or on its protein product.

  This ends the very long Cross-References section for entry P00533. If you've been clicking many of these links, you've pretty much found out everything there is to know about your protein.

The Keywords

The following section of the entry is the Keywords field (on the top of Figure 4-4). It is a simple list of terms relevant to your current protein, such as transmembrane, receptor, glycoprotein, and so on. Clicking any one of these keywords brings out a list of all Swiss-Prot entries that contain the same term. With the increasing size of the database, we're not sure that this type of query is truly useful anymore. The keyword *receptor,* for example, retrieved more than 1,570 entries!

Using the Advanced search and Full text-search features of the ExPASy server's SRS search (see Chapter 2) allows you to query Swiss-Prot by using keywords in combination. This technique gives you more specific searches.

The Features

Farther down the P00533 entry, we enter the Features section, where information on the protein is precisely mapped onto the sequence. This concept will be familiar if you've worked with GenBank (as described in Chapter 3) — fortunately, it's much simpler for proteins.

The main part of Figure 4-4 shows the beginning of the Features section for P00533.

The whole section is organized in a table with the following self-explanatory headings:

```
Key   From   To   Length   Description
```

The `Key` field indicates the kind of feature you're looking at. The numbers in the `From` and `To` fields correspond to the N-terminus-through-C-terminus numbering of the sequence, whereas the `Length` field does the math for you and gives the length of the sequence. The `Description` field gives you a brief written summary of the feature.

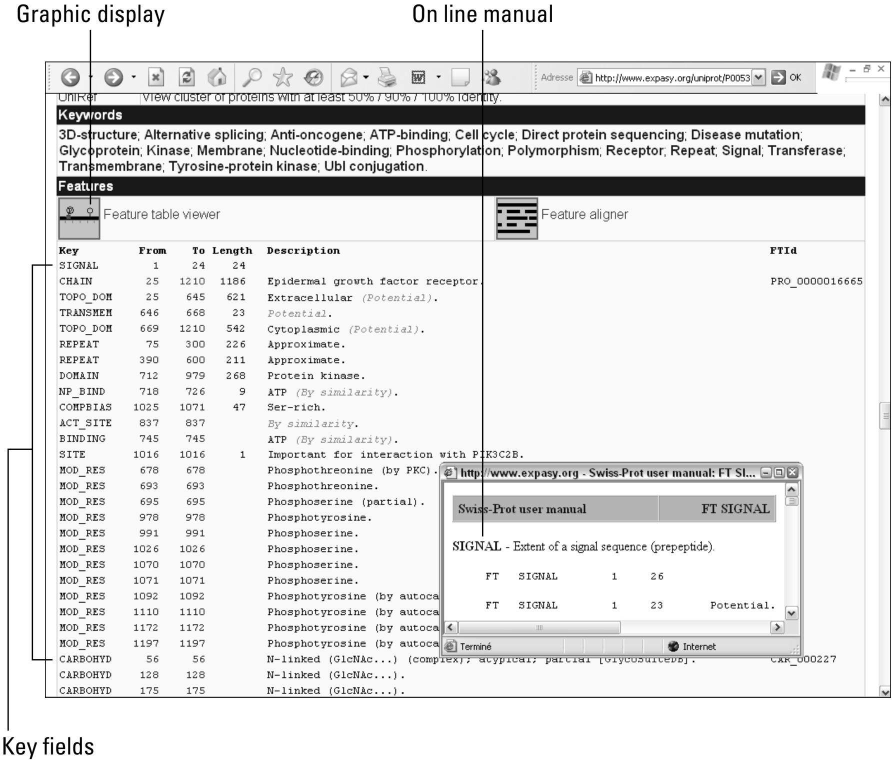
Going down the `Key` field column for the P00533 entry (refer to Figure 4-4), we can see numbers associated with the following key terms:

- **SIGNAL:** The numbers associated with this Key term indicate that residues 1 to 24 are a signal peptide, as expected for a transmembrane receptor protein.

  Notice that clicking the SIGNAL link (like any other keyword link in this field) brings out the relevant part of the Swiss-Prot user manual (refer to Figure 4-4) where you can read about the precise meaning of this term. Clicking the underlined sequence range (for example, 1 24) highlights this segment in the display of the sequence.

- **CHAIN:** Here we see that the mature peptidic chain goes from residue 25 to 1210 (the C-terminus), meaning that the signal peptide is indeed excised from the final protein.
- **TOPO_DOM:** Topological domain. Indicates which parts of the sequence correspond to what is exposed outside of the cell (the extracellular domain) or remains inside (the intracellular domain).
- **TRANSMEM:** Here we see that the following 23 residues (646–668) are the transmembrane segment of the protein.
- **DOMAIN:** These numbers indicate another domain. Residues 669 to 1210 are all inside the cell, forming the intracellular domain of the protein.

  Notice that what Swiss-Prot calls DOMAIN here, corresponds to a broader topological definition than what specialists usually mean with this word in databases like InterPro or Pfam. For instance, the extracellular domain of our EGFR protein contains several InterPro domains. (Click the links to see for yourself.)



UniRef View cluster of proteins with at least 50% / 90% / 100% identity.

Keywords

3D-structure; Alternative splicing; Anti-oncogene; ATP-binding; Cell cycle; Direct protein sequencing; Disease mutation; Glycoprotein; Kinase; Membrane; Nucleotide-binding; Phosphorylation; Polymorphism; Receptor; Repeat; Signal; Transferase; Transmembrane; Tyrosine-protein kinase; Ubl conjugation.

Features

Feature table viewer   Feature aligner

```
Key         From    To Length  Description                                                FTId
SIGNAL         1    24    24
CHAIN         25  1210  1186   Epidermal growth factor receptor.                          PRO_0000016665
TOPO_DOM      25   645   621   Extracellular (Potential).
TRANSMEM     646   668    23   Potential.
TOPO_DOM     669  1210   542   Cytoplasmic (Potential).
REPEAT        75   300   226   Approximate.
REPEAT       390   600   211   Approximate.
DOMAIN       712   979   268   Protein kinase.
NP_BIND      718   726     9   ATP (By similarity).
COMPBIAS    1025  1071    47   Ser-rich.
ACT_SITE     837   837         By similarity.
BINDING      745   745         ATP (By similarity).
SITE        1016  1016     1   Important for interaction with PIK3C2B.
MOD_RES      678   678         Phosphothreonine (by PKC).
MOD_RES      693   693         Phosphothreonine.
MOD_RES      695   695         Phosphoserine (partial).
MOD_RES      978   978         Phosphotyrosine.
MOD_RES      991   991         Phosphoserine.
MOD_RES     1026  1026         Phosphoserine.
MOD_RES     1070  1070         Phosphoserine.
MOD_RES     1071  1071         Phosphoserine.
MOD_RES     1092  1092         Phosphotyrosine (by autoca
MOD_RES     1110  1110         Phosphotyrosine (by autoca
MOD_RES     1172  1172         Phosphotyrosine (by autoca
MOD_RES     1197  1197         Phosphotyrosine (by autoca
CARBOHYD      56    56         N-linked (GlcNAc...) (complex); atypical; partial [GlycoSuiteDB].   CAR_000227
CARBOHYD     128   128         N-linked (GlcNAc...).
CARBOHYD     175   175         N-linked (GlcNAc...).
```

http://www.expasy.org - Swiss-Prot user manual: FT SI...

Swiss-Prot user manual	FT SIGNAL

SIGNAL - Extent of a signal sequence (prepeptide).

```
    FT   SIGNAL        1     26

    FT   SIGNAL        1     23        Potential.
```

Terminé   Internet

Figure 4-4: Top of the Features section for Swiss-Prot Entry P00533.

We find the DOMAIN keyword used again to indicate a serine-rich segment, and the region of the sequence responsible for the protein kinase activity. This confirms what we've just said about the loose usage of this term in Swiss-Prot. Let's keep moving, however, in search of newer terms:

- **NP-BIND:** The numbers here indicate the extent of the nucleotide phosphate binding region — from residue 718 to 726 — for binding ATP.
- **BINDING:** Here we can see the precise binding site for ATP — on lysine at position 745.
- **ACT_SITE:** The numbers here indicate amino acids involved in the activity of an enzyme.

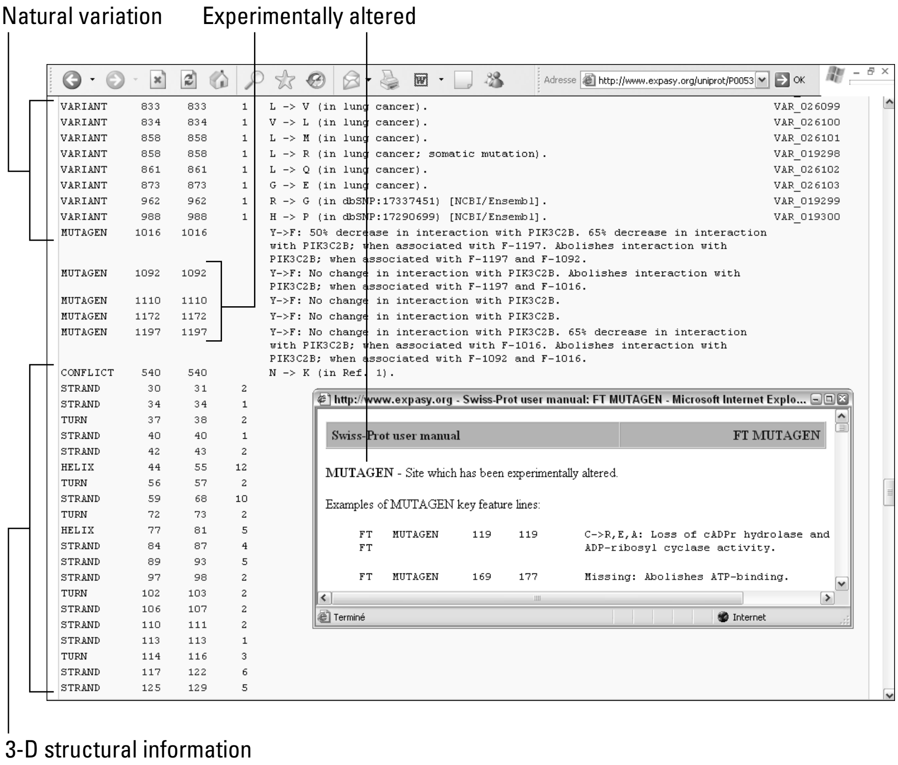
  To the nonspecialist, the three Key field terms listed above address fairly overlapping concepts.
- **COMPBIAS:** Extent of a compositionally biased (in this case, a serine rich) region of the protein sequence.

Next comes some information about residues that are chemically modified after translation (as shown in Figure 4-5). The corresponding keywords are:

- **MOD_RES:** The numbers here indicate residues undergoing phosphorylation. Click MOD_RES to find out the list of chemical modifications associated with that keyword.
- **CARBOHYD:** Here we can see the numerous residues on which carbohydrate molecules have been attached as well as the type of link (C–, N–, or O–). Note that they're all in the extracellular domain, as one would expect. Clicking the CARBOHYD keyword can tell you much more about the intricacies of glycosylation.
- **DISULPHID:** The numbers here indicate cysteine pairs forming a bond in the mature protein; there are plenty such pairs here.

Finally, the Features section ends up with a series of specialized features:

- **VAR_SEQ:** This field is used to indicate amino-acid changes from the translation of different mRNAs (isoforms) generated by alternative splicing.
- **VARIANT:** This field identifies natural variation of the EGFR protein sequence, most of which have been associated with lung cancer.
- **MUTAGEN:** In contrast with the previous field, this field is used to record sequence changes that have been experimentally (voluntarily) introduced in the protein.
- **CONFLICT:** This feature is used to indicate discrepancies between different sources of the same protein sequence — basically errors or unrecognized polymorphisms.


Natural variation
Experimentally altered
3-D structural information
Adresse http://www.expasy.org/uniprot/P0053 OK

```
VARIANT     833    833     1    L -> V (in lung cancer).                                VAR_026099
VARIANT     834    834     1    V -> L (in lung cancer).                                VAR_026100
VARIANT     858    858     1    L -> M (in lung cancer).                                VAR_026101
VARIANT     858    858     1    L -> R (in lung cancer; somatic mutation).              VAR_019298
VARIANT     861    861     1    L -> Q (in lung cancer).                                VAR_026102
VARIANT     873    873     1    G -> E (in lung cancer).                                VAR_026103
VARIANT     962    962     1    R -> G (in dbSNP:17337451) [NCBI/Ensembl].              VAR_019299
VARIANT     988    988     1    H -> P (in dbSNP:17290699) [NCBI/Ensembl].              VAR_019300
MUTAGEN    1016   1016          Y->F: 50% decrease in interaction with PIK3C2B. 65% decrease in interaction
                                with PIK3C2B; when associated with F-1197. Abolishes interaction with
                                PIK3C2B; when associated with F-1197 and F-1092.
MUTAGEN    1092   1092          Y->F: No change in interaction with PIK3C2B. Abolishes interaction with
                                PIK3C2B; when associated with F-1197 and F-1016.
MUTAGEN    1110   1110          Y->F: No change in interaction with PIK3C2B.
MUTAGEN    1172   1172          Y->F: No change in interaction with PIK3C2B.
MUTAGEN    1197   1197          Y->F: No change in interaction with PIK3C2B. 65% decrease in interaction
                                with PIK3C2B; when associated with F-1016. Abolishes interaction with
                                PIK3C2B; when associated with F-1092 and F-1016.
CONFLICT    540    540          N -> K (in Ref. 1).
STRAND       30     31     2
STRAND       34     34     1
TURN         37     38     2
STRAND       40     40     1
STRAND       42     43     2
HELIX        44     55    12
TURN         56     57     2
STRAND       59     68    10
TURN         72     73     2
HELIX        77     81     5
STRAND       84     87     4
STRAND       89     93     5
STRAND       97     98     2
TURN        102    103     2
STRAND      106    107     2
STRAND      110    111     2
STRAND      113    113     1
TURN        114    116     3
STRAND      117    122     6
STRAND      125    129     5
```

http://www.expasy.org - Swiss-Prot user manual: FT MUTAGEN - Microsoft Internet Explo...

Swiss-Prot user manual	FT MUTAGEN

MUTAGEN - Site which has been experimentally altered.

Examples of MUTAGEN key feature lines:

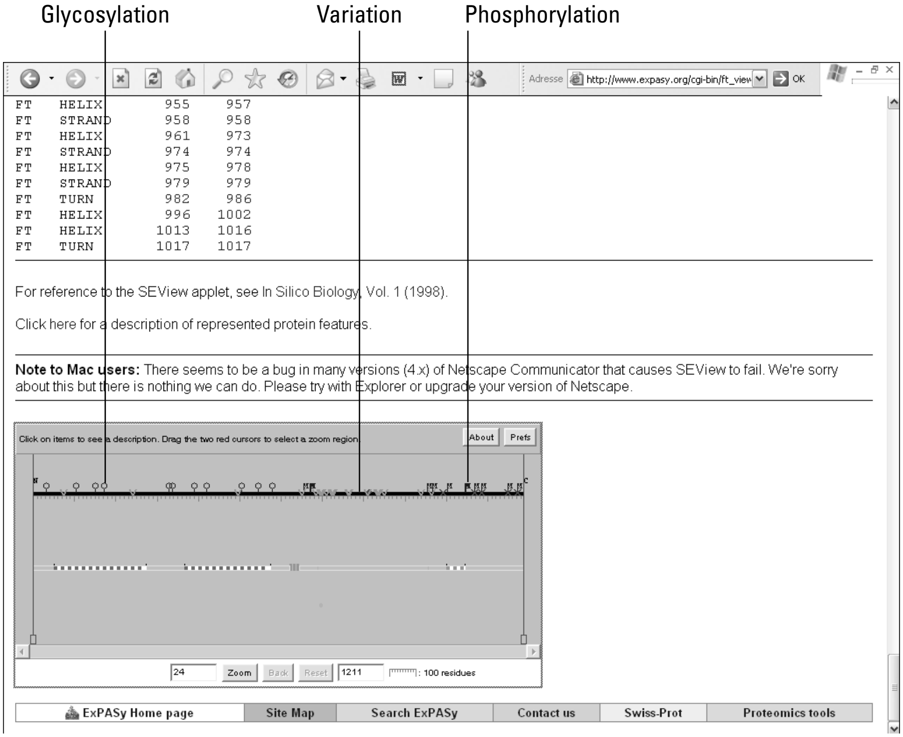
```
    FT   MUTAGEN     119    119       C->R,E,A: Loss of cADPr hydrolase and
    FT                                ADP-ribosyl cyclase activity.

    FT   MUTAGEN     169    177       Missing: Abolishes ATP-binding.
```

Terminé Internet

Figure 4-5: The bottom of the Features section for entry P00533.

Following these features lines, we enter into a detailed linear description of the local shapes (called the *secondary structure*) of the backbone of the protein. Secondary structures come in three flavors: STRAND (an extended, stringy shape), TURN (as it says!), and HELIX (a twisted, springlike shape).

Before leaving this long Features section, look along the left side for the Feature Table Viewer icon (refer to Figure 4-4, top left). Click the icon and scroll down the new page that appears to see a little graphic display, as shown in Figure 4-6. It gives you a global picture of the EGFR protein primary structure. Mousing over this picture tells you what feature corresponds to each symbol. With P00533, this shows you at a glance that all the phosphorylation events take place in the intracellular domain — while all the other modifications (disulphide bridges and glycosylations) occur in the extracellular domain. If you've done a bit of biochemistry, you know that this makes sense and is as it should be!



```
FT   HELIX      955    957
FT   STRAND     958    958
FT   HELIX      961    973
FT   STRAND     974    974
FT   HELIX      975    978
FT   STRAND     979    979
FT   TURN       982    986
FT   HELIX      996   1002
FT   HELIX     1013   1016
FT   TURN      1017   1017
```

For reference to the SEView applet, see In Silico Biology, Vol. 1 (1998).

Click here for a description of represented protein features.

Note to Mac users: There seems to be a bug in many versions (4.x) of Netscape Communicator that causes SEView to fail. We're sorry about this but there is nothing we can do. Please try with Explorer or upgrade your version of Netscape.

Click on items to see a description. Drag the two red cursors to select a zoom region.

Figure 4-6: The Feature table viewer for Swiss-Prot entry P00533.

Finally, the sequence itself

There's nothing special to say about this straightforward section. Remember that you can get the sequence in the more convenient FASTA format version by clicking the P00533 in FASTA format link, located at the bottom right of the entry. Use the File➪Save As option of your browser to save it, or cut and paste it into a word processor.

Finding Out More about Your Protein

After carefully reading the Swiss-Prot entry on your protein of interest (or its homologue) and using all the hyperlinks it provides, you probably know 90 percent of what the whole world knows about this protein.

Your next problem is to make sense of, integrate, or eventually understand what you have read, in Swiss-Prot or through its numerous hyperlinks. For instance, perhaps you're not quite sure what myristilation is? Or you don't necessarily understand the biochemical reaction catalyzed by a protein? Or you'd like to know in which pathway this enzyme is involved?

In brief, there may come a time when you need additional information to better understand what you've read in Swiss-Prot. In the following sections, we point out some extremely valuable resources for our knowledge-craving readers!

Finding out more about modified amino acids

If you want to find out everything you always wanted to know about the modification of amino acids during protein maturation, have a look at RESID(r), the post-translational modification database maintained by John Garavelli at the European Bioinformatics Institute. For instance, you can use this resource to quickly find out what myristylation is exactly. Here's how:

1. **Point your browser to** `www.ebi.ac.uk/RESID/`.

   The RESID Database of Protein Modifications page appears.

2. **Click the Search RESID link.**

   A search form appears.

3. **Enter** myristoylation **in the search window (refer to Figure 4-7).**

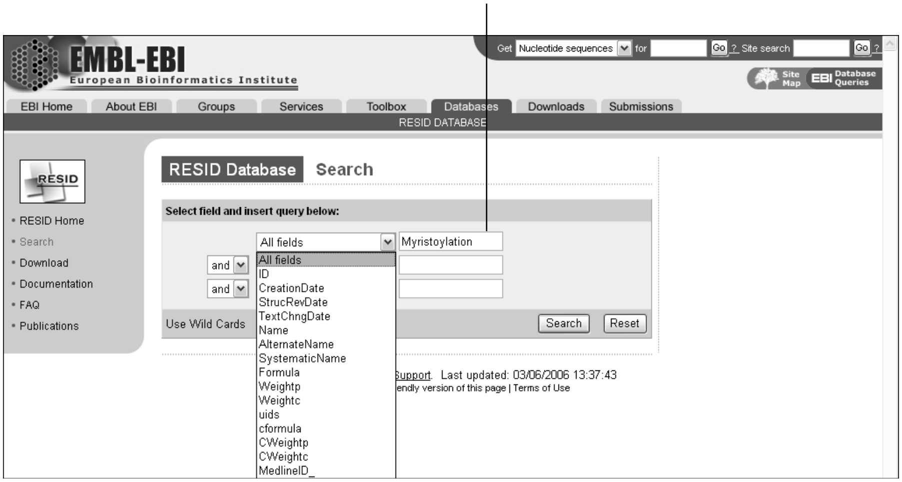
4. **Click the Search button.**

   The query returns a table with 3 entries: AA0059 (N-myristoyl-glycine), AA0078 (N6-myristoyl-L-lysine), and AA0307 (S-myristoyl-L-cysteine)

5. **Click the AA0078 link in the RESID ID leftmost column.**

   You obtain a complete report on this modified amino acid, ending with its detailed chemical formula.

Not impressed with this one? Why not have a go at the modified cysteine — L-cysteinyl molybdopterin guanine dinucleotide (AA0281)? Now, *there's* a compound name to conjure with; reading it aloud several times can only help the processing of your query!



Enter search term.

Figure 4-7: Querying the RESID database for myristoylation.

Some advanced biochemistry sites

If you want to brush up your biochemistry and chemistry skills, visiting the following sites could help:

- **The Glycan Structure Database at** `www.glycosuite.com/`

  This one is free for academics only — and registration is required.

- **The Lipid Bank at** `lipidbank.jp`
- **ChemIDplus at** `chem.sis.nlm.nih.gov/chemidplus/`

  This one is fun; you can query the databank by drawing the molecule of your dreams!

WARNING!

Be aware that using these databases is tricky for nonspecialists. We only introduce them here so you know they exist. There is some brushing up on biochemical paths to do before you can search them in a sensible way.

Finding out more about biochemical pathways

Reading that your protein is involved in the biosynthesis of di-amino-pymelate probably rings no bell at all! To find out which biochemical pathway (a succession of elementary reactions leading to a compound central to the cell's well being) your protein belongs to, pay a visit to the Boehringer site.

If you've ever done any formal biochemical studies, we're pretty sure you've seen this huge chart hanging on a wall in your biochemistry department. Did you know that you can search it automatically these days? Yep, this is the kind of era we're living in!

1. **Point your browser to** `www.expasy.ch/cgi-bin/search-biochem-index`**.**
2. **Type** pimelate **in the Keyword Search window and then click the Submit Query button.**

   Your browser displays the Pimelate page. Pimelate — a central molecule for the creation of bacterial cell walls — is just an example here. If you like, you can substitute the name of the compound you're working on at the moment.
3. **Click J3 below** `6-DIAMINOPIMELATE`**, the first entry in the list.**

   This brings up a local chart where you can see that this compound is central to the biosynthesis of bacterial cell walls.

Table 4-1 lists more great biochemical-pathway resources.

Table 4-1 Biochemical Pathways Resources Available Online

Address	*Description*
`www.genome.ad.jp/kegg/`	The famous Kyoto Encyclopedia of Genes and Genomes (KEGG). E.C. numbers or gene names are the best starting points for this resource.
`brenda.bc.uni-koeln.de`	The Comprehensive Enzyme Information System BRENDA is a must if you plan to get into real experiments!
`www.chem.qmul.ac.uk/iubmb/`	The official site for Enzyme Nomenclature of the International Union of Biochemistry and Molecular Biology (IUBMB).
`www.ecocyc.org`	The Encyclopedia of E. coli Genes and Metabolism. Now extending progressively to other bacteria.

Finding out more about protein structures

If you have analyzed your protein at the sequence level by all available means (motif search in Chapter 6, database search in Chapter 7, multiple alignment in Chapter 9), your next question probably relates to the position of certain amino-acid residues in your protein 3-D structure. Is this amino acid buried or at the surface? Is it close or far from another residue? Is it in the active site? The list of questions goes on. All these questions require structural information. In Chapter 11, we show you how to relate sequences and structural information. For now, Table 4-2 gives you a list of the major resources dedicated to this topic.

Table 4-2 Databases Dedicated to Structural Information

Address	*Description*
`www.rcsb.org/pdb/`	PDB, the official repository database for protein 3-D structures.
`www.ncbi.nlm.nih.gov/Structure/`	with tools for their visualization and comparative analysis.
`scop.mrc-lmb.cam.ac.uk/scop/`	SCOP, a Structural Classification Of Proteins, aims to provide descriptions of the structural and evolutionary relationships among all proteins whose structure is known.
`www.cathdb.info`	CATH (*C*lass, *A*rchitecture, *T*opology, *H*omologous superfamily) also provides a hierarchical classification of protein-domain structures.
`swissmodel.expasy.org`	Swiss-Model is a fully automated server that models protein-structure homology, accessible via the ExPASy server.

Finding out more about major protein families

Some protein families are so hot they've become whole worlds of their own. There is always a time when general protein databases just can't manage the sheer amount of detailed information required by the specialized research community. Highly specialized information resources exist to fulfill that niche. If your favorite protein belongs to one of the following categories, you may want to try some of the resources that we list in Table 4-3.

Table 4-3 Some Specialized Protein Databases

Address	*Description*
`imgt.cines.fr`	IMGT, the International Immunogenetics database, specializes in Immunoglobulins, t cell receptors, and major histocompatibility complex molecules of all vertebrate species.
`rebase.neb.com`	REBASE is a restriction/modification enzyme database.
`www.cazy.org/CAZY/`	CAZy is an information resource on enzymes that degrade, modify, or create glycosidic bonds.
`merops.sanger.ac.uk`	MEROPS is a database providing a wealth of information on proteases.
`www.kinasenet.org/pkr/`	PKR, the Protein Kinase Resource, is a Web-accessible compendium of information on the protein kinase family of enzymes.
`www.nursa.org`	Nursa, the Nuclear Receptor Signaling Atlas, is a great information resource on nuclear (for example, steroid) receptors together with their coactivators, corepressors, and ligands.
`senselab.med.yale.edu/senselab`	The Human Brain Database provides information on the proteins involved in neural processes, such as ion channels, membrane receptors of neurotransmitters and neuromodulators, and olfactory receptors.
`www.ncbi.nlm.nih.gov/COG`	The COG (Clusters of Orthologous Groups) database regroups proteins shared by at least three major phylogenetic lineages, thus corresponding to ancient conserved domains.

Chapter 5

Working with a Single DNA Sequence

In This Chapter

- Identifying problems with your sequence
- Computing and displaying a restriction map
- Profiling your sequence for a variety of properties
- Finding ORFs, exons, and genes
- Assembling sequence fragments

> *Not everything that can be counted counts, and not everything that counts can be counted.*
>
> — Albert Einstein (1879–1955)

Proteins perform most biological functions, so biologists tend to consider DNA sequences kind of dull. They're mostly right. The truth of the matter is, 90 percent of the DNA you may use in your work is nothing more than an information-storing device, sort of the organic equivalent of a floppy disk or a DVD. Still, we probably shouldn't turn up our noses at information-storing devices. DNA contains information that the cell knows how to put to effective use, as quickly as your computer can turn a dull-looking DVD into the latest destroy-all-at-any-cost network game.

If you're interested in a protein, directly working out its amino-acid sequence is hell. Sequencing its gene and then deducing the protein sequence from the genetic code is much easier. This is why DNA sequences are so important. These days, working with DNA sequences is an obligatory intermediate step; you simply must know how to handle a nucleotide sequence in order to work with protein or RNA sequences. DNA sequences are the mothers of all sequences!

If you're reading this book, it's probably fair for us to assume that you're not a highly trained scientist working in one of those mega-super-duper genome-sequencing centers churning out nucleotides by the millions every day. This chapter is thus focused on the analysis of relatively small sequences — 100,000 nucleotides long at most. The skills necessary for dealing with genomic fragments of a reasonable size or with complementary DNA — cDNA, the DNA copy of a messenger RNA — is the name of the game in this chapter. This is the kind of sequence you can routinely determine in a well-equipped molecular biology lab or get done for a reasonable price by a sequencing company.

When you get an experimental sequence back, your first task is to make sure it's okay so you don't waste a month trying to make sense of something completely botched and end up nowhere. This chapter begins with a section showing you how to check for the most common problems encountered in sequences fresh from the bench. But remember Murphy's Law: If something can go wrong, *and has not been discussed in this section,* it will happen to you!

Catching Errors Before It's Too Late

Sequencing a DNA fragment involves purifying it, cloning it into a vector (such as a plasmid), amplifying it into a biological host (most often a bacterium like *E. coli*), and finally submitting it to one of the various sequencing protocols, such as primer extension or dye-termination. During this process, accessory DNA segments are deliberately linked to your target DNA. In addition, many unintended events can occur, resulting in a sequence that doesn't faithfully correspond to the genetic information you intended to study.

The most common problem is that the sequence you get back has been contaminated with some of the sequence of the vector you used (or that somebody else used in the laboratory). Recognizing this problem — and knowing how to deal with it — is so important that we devote the next section to precisely this topic.

Removing vector sequences

The DNA (or cDNA) you send out for sequencing is inserted into a cloning vector — plasmid, phage, cosmid, BAC, PAC, or YAC — so that it can be manipulated. The sequences you get from such constructs inevitably include segments derived from these vectors, at least at one extremity or the other. You can detect and remove them from your sequence by simply running a search for similarity against the sequence of the vector you have been using. (As a responsible scientist, you're expected to have this information!)

However, your sequence might also be cross-contaminated by somebody else's vector — maybe from the bench next to yours. It's thus wiser to search your sequence not only for the vector you expect, but also for other possible vector contaminations, just in case. Here's how to do it using a nice interactive tool available at the NCBI (the *N*ational *C*enter of *B*iotechnology *I*nformation):

1. **Point your browser to** `www.ncbi.nlm.nih.gov/VecScreen/VecScreen_docs.html`**.**

   The entry page for VecScreen appears. It contains both a nice tutorial on sequence contamination (click the Contamination link to read it) as well as an explanation of how VecScreen works. Basically, it performs a blastn search (see Chapter 7) of your sequence against a special database of all vectors and cloning adapter sequences known to man. This database is known as UniVec, and a simple click the UniVec link on this page provides you with more details about this unique database.

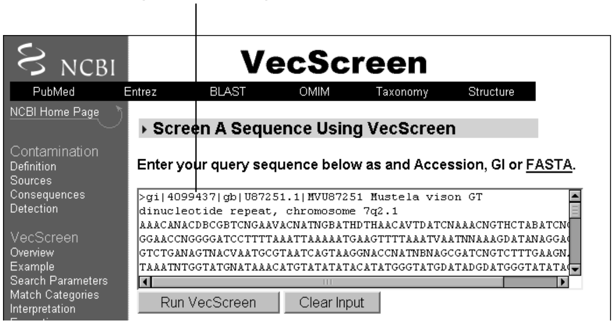
   When you believe you're ready for the real thing, rather than just reading about the real thing, go grab your new sequence and proceed to Step 2.

2. **Point your browser to** `www.ncbi.nlm.nih.gov/VecScreen/`**.**

   A typical sequence-input window appears, ready for your input.

3. **Paste in your newly determined sequence in the input window, as shown in Figure 5-1.**

   You can use FASTA or raw format.


Paste in your new sequence.

Screen A Sequence Using VecScreen

Enter your query sequence below as and Accession, GI or FASTA.

```
>gi|4099437|gb|U87251.1|MVU87251 Mustela vison GT
dinucleotide repeat, chromosome 7q2.1
AAACANACDBCGBTCNGAAVACNATNGBATHDTHAACAVTDATCNAAACNGTHCTABATCN
GGAACCNGGGGATCCTTTTAAATTAAAAATGAAGTTTTAAATVAATNNAAAGDATANAGGA
GTCTGANAGTNACVAATGCGTAATCAGTAAGGNACCNATNBNAGCGATCNGTCTTTGAAGN
TAAATNTGGTATGNATAAACATGTATATATACATATGGGTATGDATADGDATGGGTATATA
```

Figure 5-1: The VecScreen input window.

4. **Click the Run VecScreen button below the input window.**

   You immediately get back the BLAST formatting page, showing you that your VecScreen request was in fact a BLAST search in disguise. (For more on BLAST, see Chapter 7.)

It's better to use VecScreen from the start rather than trying to use blastn because the VecScreen server chooses parameter values that are best for vector detection.

5. **Click the Format! button and wait.**

   Don't change any of the default parameters.

   At this point — and depending on the workload of the BLAST server — you'll either get the standard waiting page (`This page will be automatically updated in` *`xx`* `seconds`), or the results of your search pop up immediately.

6. **The two possible outcomes are**

   - **Non-significant similarity found.**

     This is good news! This message indicates that the sequence you submitted does not resemble any known vector/adapter. You can proceed to the next stage of its analysis, following the rest of this chapter.

   - **An output listing matches of some kind.**

     This is bad news! Your query does contain some vector/query sequence.

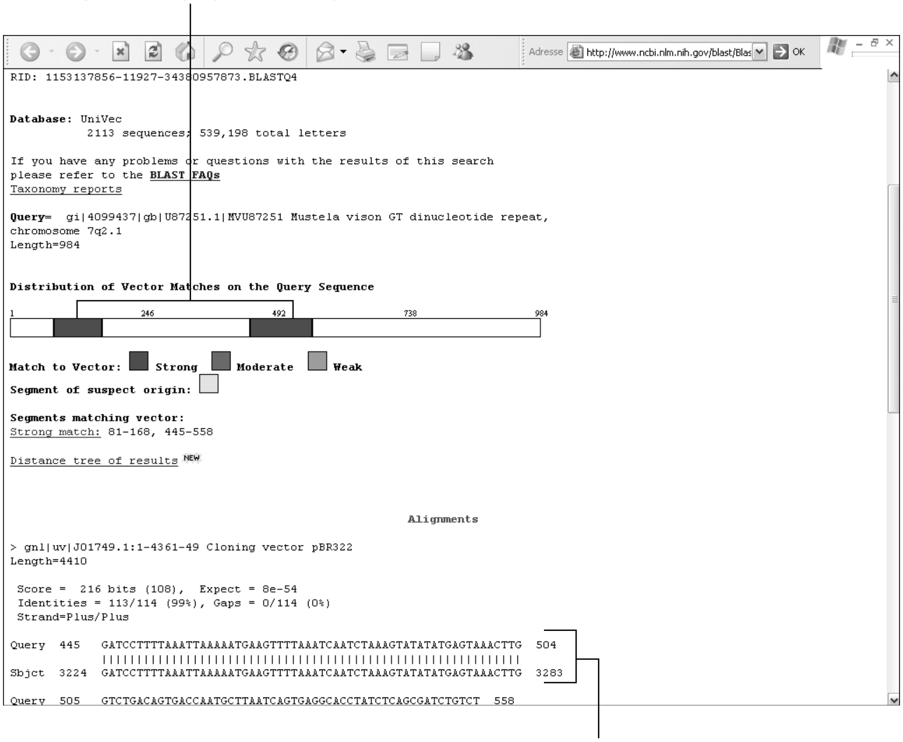
A VecScreen contamination report output contains two parts. The top part displays the distribution of the contaminating sequences along the length of your query sequence, using a color-coded graph (Figure 5-2 — you'll have to go there online to see it in color). The second part is a standard blastn output, providing you with the identity of the matched vector/adaptor as well as the detailed alignment.

The different colors on the graph correspond to the strong (red), moderate (purple), or weak (green) matches in the UniVec database. A yellow code is used to delineate segments of suspect origin such as short segments in between two vector matches. The U87251 exhibits two regions of 100 percent identity with the popular PBR322 cloning vector. Most likely, it is a contaminated sequence that was submitted to GenBank before vector-sequence filtering was systematically included in the submission process.

What to do next with contaminated sequences is fully explained in a complete tutorial that you can follow by clicking the Interpretation of VecScreen Results link provided at the top of any VecScreen positive output.

In general, you have two options when finding contaminations. If the contaminated parts are at the extremity of your sequence and correspond to the vector you used, all is fine. Just remove these parts and proceed. If, on the other hand, contaminated parts are found elsewhere (as in Figure 5-2), or if the contaminating vector is not yours, you'd better consider this sequence to be the result of a chimeric clone or a DNA cross-contamination. In that case, throwing it away is the safest thing to do!

Your sequence displays two strong vector contaminants.



```
RID: 1153137856-11927-34380957873.BLASTQ4


Database: UniVec
           2113 sequences; 539,198 total letters

If you have any problems or questions with the results of this search
please refer to the BLAST FAQs
Taxonomy reports

Query=  gi|4099437|gb|U87251.1|MVU87251 Mustela vison GT dinucleotide repeat,
chromosome 7q2.1
Length=984


Distribution of Vector Matches on the Query Sequence

1                  246                  492                  738                  984

Match to Vector: ■ Strong  ■ Moderate  ■ Weak
Segment of suspect origin: □

Segments matching vector:
Strong match: 81-168, 445-558

Distance tree of results NEW



                                                         Alignments

> gnl|uv|J01749.1:1-4361-49 Cloning vector pBR322
Length=4410

 Score =  216 bits (108),  Expect = 8e-54
 Identities = 113/114 (99%), Gaps = 0/114 (0%)
 Strand=Plus/Plus

Query  445   GATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTG  504
             ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
Sbjct  3224  GATCCTTTTAAATTAAAAATGAAGTTTTAAATCAATCTAAAGTATATATGAGTAAACTTG  3283

Query  505   GTCTGACAGTGACCAATGCTTAATCAGTGAGGCACCTATCTCAGCGATCTGTCT  558
```

Blast N sequence alignment

Figure 5-2: VecScreen output display of a contaminated sequence.

Cases when you shouldn't discard your sequence

In case of a VecScreen match, do not merely rely on the name of the UniVec database entry to decide that your experiment was contaminated by somebody else's vector. You may have used pUC19 as a cloning vector and see that VecScreen is reporting a strong match with pUR222 or lactose operon genes from *E. coli.* This is okay. Most commercial vectors are derived from the same initial natural plasmid and *E. coli* constructs. Their sequences must thus be 100 percent identical. The UniVec matches are simply reported in the order that they occur in the database. If your vector is not first, another entry (corresponding to a locally 100-percent-identical sequence) will be reported instead.

Also, if you're working on genes that resemble the kinds of genes used to build vectors (such as antibiotic resistance genes, repressor/activator genes, or some biosynthetic genes), you should actually *expect* matches in the UniVec database.

Don’t throw away these sequences before you interpret the detailed alignment.

In any case, beware of near 100-percent matches with any piece of the *E. coli* or yeast genome, even if these parts don’t correspond to vectors. These are sure signs of cross-contamination. Modern molecular biology techniques are very good at amplifying pieces of DNA from the bench next to yours!

Computing/Verifying a Restriction Map

Here’s another way of checking whether the sequence you get back from the sequencing lab or company actually corresponds to the piece of DNA you sent out: Compare the pattern of restriction enzyme digestion (also called a *restriction map*) that you experimentally observe with the pattern (*theoretical* restriction map) that you can compute from the sequence.

Computing a theoretical restriction map is easy. All you do is look for exact matches of a given restriction-enzyme site within your sequence. All the known enzymes and sites are described in the REBASE database (`rebase.neb.com`). After locating the matches in the sequence, it’s simply a matter of counting the number of resulting fragments (DNA segments between two successive sites) and computing their sizes. You can then compare the computer prediction with your lab experiment. The more restriction enzymes you test, the more certain you’ll be that your sequence is correct.

You can use this approach to verify long sequences — such as bacterial genome sequences — that were assembled from many shorter ones, as happens with the Shotgun DNA sequencing protocol. If the sequence and its assembly are correct, the number and size of the predicted fragments should perfectly correspond to the experimental restriction map.

A couple of good Internet sites spare you all the work of having to collect the information about restriction-enzyme sites and doing the matching yourself. These sites maintain their own current versions of the authoritative Restriction Enzyme Database, REBASE, so you don’t have to worry about it. They scan your sequence for all the specific sites corresponding to an enzyme list that you can select according to various criteria.

One such Internet site we find especially useful offers a nice little interactive tool called Webcutter, developed by Max Heiman. Here’s how to put it to work for you:

1. **Point your browser to** `www.firstmarket.com/cutter/cut2.html`**.**

   This gives you access to the most current version of Webcutter.

2. **Enter the sequence you'd like to check.**

   At the time of this writing, Webcutter offers you three ways to input your sequence:

   - Copy and paste it into the Sequence window displayed on the `www.firstmarket.com/cutter/cut2.html` page.
   - Upload the sequence from your computer (Netscape users only).

     WARNING!

     The sequence has to be in a text format, without its name or any FASTA-style header.
   - Directly fetch a GenBank entry by using its accession number or other identification methods. (The latter case applies if you're designing a cloning experiment for a gene already known.)

3. **Select the options that correspond to your problem.**

   Options include

   - Choosing whether to treat the input sequence as linear or circular
   - Selecting various output formats for the results (maps, table of enzymes)
   - Filtering the results based on whether enzymes cut or don't cut your sequence — or cut your sequence a given number of times
   - Specifying a subset of enzymes — based on various criteria — to be included in the analysis

4. **Click the Analyze Sequence button at the bottom of the form.**

   The result usually arrives very quickly and is — thankfully — self-explanatory.

The University of Massachusetts Medical School offers a similar service with a different interface at `biotools.umassmed.edu/tacg/WWWtacg.php`.

TIP

For a list of other restriction mapping servers, check out the REBASE Related Sites link on the REBASE home page (`rebase.neb.com`). All these servers are equivalent in terms of accuracy. Shop around until you find the one whose input/output format you like best.

Designing PCR Primers

The polymerase chain reaction (PCR) is a very common experimental laboratory technique people use to amplify segments of DNA they find interesting. To make a PCR, you must

What is PCR?

Running a polymerase chain reaction (PCR) involves mixing a DNA template (containing the sequence that you mean to amplify), the primers, a cocktail of nucleotides and other biochemical compounds, and a heat resistant enzyme called DNA polymerase — all in a single little plastic tube. You then put that tube in a benchtop expensive machine called a *thermal cycler.* According to a preset program, this machine will make the tube go up and down in temperature. For each temperature cycle, the amount of copies of the DNA segment that you want to amplify is doubled. For instance, after 30 cycles, you have 2^{30} (1 billion) more of it. This is why PCR is such a great technique in forensic science: Lick a stamp, and scientists will be able to sequence your DNA!

If you want to know more about PCR, visit `nature.umesci.maine.edu/forensics/p_intro.htm` for a great animated introductory course.

1. **Identify the DNA sequence you want to amplify.**
2. **Order primers from a DNA synthesis company.**

   Primers are small pieces of DNA (20–30 bases) that match the extremities of the complete sequence you're interested in. Designing good primers is the most delicate step in a PCR.
3. **Run your PCR experiment.**

These days, people use PCR for almost anything that has to do with wet-lab sequence analysis — subcloning a gene in an expression vector, verifying its assembly and sequence, inducing or detecting mutations, or detecting its relatives in other organisms.

The trickiest step when you do a PCR is the design of the *primers* — the two small DNA fragments capable of firmly hybridizing on each side of your gene in a highly specific manner. Primer design programs are here to help you decide which portion of your large sequence makes the best primers, thus avoiding a series of potential problems too numerous to be listed here.

The Internet site of the University of Massachusetts Medical School (`biotools.umassmed.edu/`) provides a link to a very complete and easy-to-use tool for primer design. It is a Web implementation of Primer3, the well-known program developed by Steve Rosen and Helen Skaletsky. To put it to (hopefully good) use, do the following:

1. **Point your browser to `biotools.umassmed.edu/`.**

   The Bioinformatics Resources page of the University of Massachusetts Medical School appears, offering access to several sequence-analysis programs that you may want to try by yourself later.

2. **Click the Primer Selection Tool link in the DNA Sequence Analysis section.**

   TIP

   The Primer3 page dutifully appears. The form can look rather intimidating, as Figure 5-3 more than clearly shows. Don't let it scare you! In most cases, you can get by with the default settings.

3. **Paste your sequence in the Sequence window.**

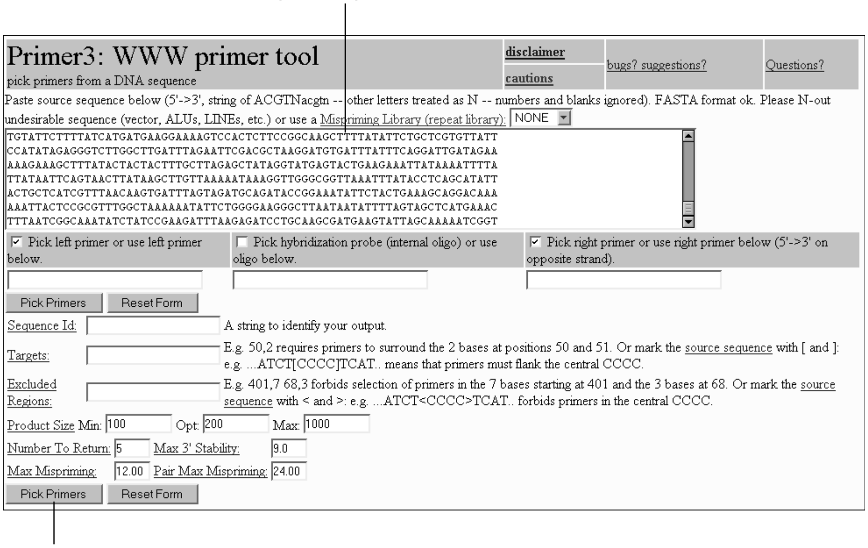
4. **Click the Pick Primers button.**

   The results return very quickly. They include a map of the best oligonucleotide pair (left and right primers) according to the constraints listed in the form. In addition, four alternative solutions (by default) are proposed. The position, length, predicted melting temperature, and G+C percentage (the fraction of guanine and cytosine nucleotides) are given for each listed oligonucleotide.

If you're ready to modify the default setting, the form allows all reasonable experimental situations and constraints to be imposed upon the primer-selection process, including

- Searching for only a left or right primer, or a single hybridization probe
- Proposing your own left or right primer
- Selecting sequence positions to be flanked or excluded
- Selecting a range of product sizes

Paste your sequence here.



Primer3: WWW primer tool

pick primers from a DNA sequence

Paste source sequence below (5'->3', string of ACGTNacgtn -- other letters treated as N -- numbers and blanks ignored). FASTA format ok. Please N-out undesirable sequence (vector, ALUs, LINEs, etc.) or use a Mispriming Library (repeat library): NONE

```
TGTATTCTTTTATCATGATGAAGGAAAAGTCCACTCTTCCGGCAAGCTTTTATATTCTGCTCGTGTTATT
CCATATAGAGGGTCTTGGCTTGATTTAGAATTCGACGCTAAGGATGTGATTTATTTCAGGATTGATAGAA
AAAGAAAGCTTTATACTACTACTTTGCTTAGAGCTATAGGTATGAGTACTGAAGAAATTATAAAATTTTA
TTATAATTCAGTAACTTATAAGCTTGTTAAAAATAAAGGTTGGGCGGTTAAATTTATACCTCAGCATATT
ACTGCTCATCGTTTAACAAGTGATTTAGTAGATGCAGATACCGGAAATATTCTACTGAAAGCAGGACAAA
AAATTACTCCGCGTTTGGCTAAAAAATATTCTGGGGAAGGGCTTAATAATATTTTAGTAGCTCATGAAAC
TTTAATCGGCAAATATCTATCCGAAGATTTAAGAGATCCTGCAAGCGATGAAGTATTAGCAAAAATCGGT
```

☑ Pick left primer or use left primer below.

□ Pick hybridization probe (internal oligo) or use oligo below.

☑ Pick right primer or use right primer below (5'->3' on opposite strand).

Sequence Id: A string to identify your output.

Targets: E.g. 50,2 requires primers to surround the 2 bases at positions 50 and 51. Or mark the source sequence with [and]: e.g. ...ATCT[CCCC]TCAT.. means that primers must flank the central CCCC.

Excluded Regions: E.g. 401,7 68,3 forbids selection of primers in the 7 bases starting at 401 and the 3 bases at 68. Or mark the source sequence with < and >: e.g. ...ATCT<CCCC>TCAT.. forbids primers in the central CCCC.

Product Size Min: 100 Opt: 200 Max: 1000

Number To Return: 5 Max 3' Stability: 9.0

Max Mispriming: 12.00 Pair Max Mispriming: 24.00

Click here to get your primers.

Figure 5-3: Primer3 input form.

- Imposing a range of oligonucleotide sizes, G+C percentages, or melting points
- And many more highly sophisticated options

After you've gotten your primers, your final responsibility is to verify that they will not hybridize anywhere — except where you *intend* them to hybridize. For instance, you want to make sure that the selected oligonucleotide sequences aren't found outside the gene you're interested in — or (worse) resemble a frequent repeat in the DNA you're going to amplify.

The techniques for avoiding these problems involve BLAST searches against the vector sequences, the relevant genomes as well as their most common repeats. You can use the NCBI BLAST server for this purpose. (For more on BLAST searches, see Chapter 7.)

Analyzing DNA Composition

After these analyses — with all the going back and forth from computer to bench they entail — you're now convinced that the sequence in your test tube is indeed the right stuff. The time has now come to analyze it! Whenever you have a new piece of a new genome, the first question people will ask you is: *What is the percentage of G+C?* Because the pairing between the guanosine (G) and cytosine (C) nucleotides is more stable than the pairing between adenosine (A) and thymidine (T) in the DNA ladder, this percentage determines how the DNA will behave in your experiments. The first level of analysis you can perform is, thus, simply to count the number of A (adenosine), G (guanosine), C (cytosine), and T (thymidine) in your sequence.

Establishing the G+C content of your sequence

Most often, people establish G+C content by using a program installed on their own computers. Many programs that to do this are commercial packages, available from an open source consortium (such as EMBOSS), or are written up by the computer science student you keep handy and pay with pizza. If you want to save the cost of the pizza, you can even do the coding yourself if you've learned the basics of one of the basic computer-programming languages out there.

However, if you want to use an online service to establish the G+C content of your sequence for you, check out the Pasteur Institute's EMBOSS server (at bioweb.Pasteur.fr/seqanal/EMBOSS/) and click geecee in the table of available EMBOSS programs (bioweb.pasteur.fr/seqanal/interfaces/geecee.html).

Counting words in DNA sequences

After establishing the G+C content, the next step in complexity is to consider your DNA text as made up of overlapping "words." For instance, consider the following DNA sequence:

```
ATGGCTGACTGACTGACTGACTGAC......
```

You can read it as

```
A, T, G, G, C, T, G, A, C, T, G, A, C, T, G, A, ......
```

where counting the various components A, C, G, and T leads to the simple *nucleotide* statistics. But you can also read the sequence as

```
AT, TG, GG, GC, CT, TG, GA, AC, CT, TG, ....
```

and compute the frequency of the 16 different *dinucleotides* (2-letter words). Then again, you can read the same sequence as made of triplets — like this —

```
ATG, TGG, GGC, GCT, CTG, TGA, GAC, ACT, CTG, ...
```

and compute the frequency of the 64 different *trinucleotides* (3-letter words), and so on.

The German-based company Genomatix offers a nice free Web site where you can compute the nucleotide, dinucleotide, and trinucleotide frequencies of your DNA sequence. Here's how it's done:

1. **Prepare your DNA sequence in a file, using a text format.**
2. **Point your browser to** www.genomatix.de/cgi-bin/tools/tools.pl**.**
3. **Select the Create Sequence Statistics button.**
4. **Click the Start Selected Task button at the bottom left of the page.**

   A Sequence Input window appears.
5. **Cut and paste your sequence into the Sequence Input window.**

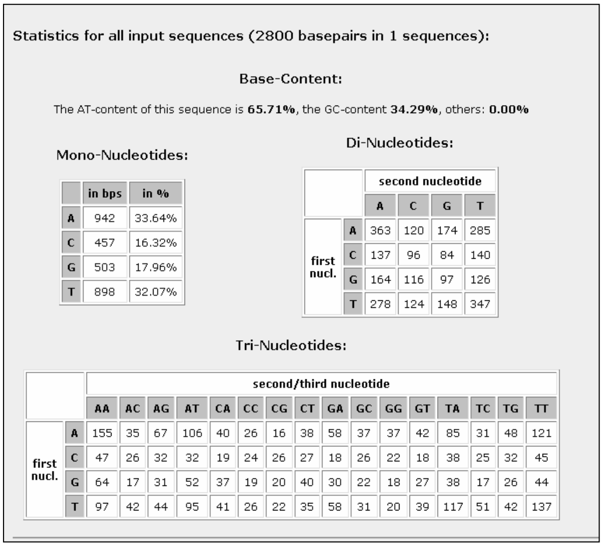
   Use the format of your choice. (Here we use the *plain* — that is, raw — format.)

6. **Click the Load Sequence button.**

   A new Input form appears.

7. **Leave the Input form as is and click the Start Task button.**

   The preset options compute the AT percent and GC percent, as well as the nucleotides, dinucleotides, and trinucleotides frequencies and presents the results, as shown in Figure 5-4.



Statistics for all input sequences (2800 basepairs in 1 sequences):

Base-Content:

The AT-content of this sequence is **65.71%**, the GC-content **34.29%**, others: **0.00%**

Mono-Nucleotides:

	in bps	in %
A	942	33.64%
C	457	16.32%
G	503	17.96%
T	898	32.07%

Di-Nucleotides:

		second nucleotide			
		A	C	G	T
first nucl.	A	363	120	174	285
	C	137	96	84	140
	G	164	116	97	126
	T	278	124	148	347

Tri-Nucleotides:

		second/third nucleotide															
		AA	AC	AG	AT	CA	CC	CG	CT	GA	GC	GG	GT	TA	TC	TG	TT
first nucl.	A	155	35	67	106	40	26	16	38	58	37	37	42	85	31	48	121
	C	47	26	32	32	19	24	26	27	18	26	22	18	38	25	32	45
	G	64	17	31	52	37	19	20	40	30	22	18	27	38	17	26	44
	T	97	42	44	95	41	26	22	35	58	31	20	39	117	51	42	137

Figure 5-4: Typical compositional analysis result.

Counting long words in DNA sequences

When analyzing DNA, it can be interesting to count words that are longer than three nucleotides, although it is rare to use words with sizes above six or eight. When their size is larger than 3, words are often referred to as k-tuples (6-letter words = hexamers = 6-tuple). There are 64 different 3-tuples, 256 4-tuples, 1,024 5-tuples, and 4,096 6-tuples. It makes no sense to compute the hexamer statistics on a short sequence where most hexamers occur, at most, only once. On the other hand, identifying hexamers (6-tuples) with unexpected high frequencies in a set of sequences (such as promoter regions) is often the starting point for discovering regulatory sequence motifs.

The EMBOSS server at the Pasteur Institute, based in Paris, offers an online version of the program wordcount that allows you to compute the word frequency in your DNA sequence for any size (we tested it up to 20 letters). Here's how you get it to work for you:

Triplet counting versus codon frequency

We introduce the concept of counting trinucleotides (triplets or 3-tuples) in this chapter. These computations are always done by reading the DNA sequence on one strand, in an overlapping manner. This should not be confused with the computation of codon usage. (Remember that a *codon* is a triplet of nucleotides used in the context of the genetic code, when translating a DNA sequence into a protein; see Chapter 1.) The concept of codon usage thus only applies to DNA sequences corresponding to protein coding regions (for example, open reading frames or ORFs). It consists in computing the frequency of trinucleotides in a non-overlapping manner, mimicking the protein translation process. Hence, for a DNA sequence such as

```
ATGTTTAGTGATGGACGCCCAGCATGC
    GAC....
```

The codon frequency is computed using the following parsing:

```
ATG, TTT, AGT, GAT, GGA, CGC,
    CCA, GCA, TGC, GAC, ...
```

The resulting table of codon usage tells you which codons are preferentially used to code for a given amino acid — and, more importantly, to determine which ones are rare. Codon usage varies among species. For instance, the codon usage in human genes is quite different from the one used by *E. coli.* For this reason, some human genes don't work well in these bacteria. Take a virtual trip to Japan at

```
www.kazusa.or.jp/codon/
    countcodon.html
```

to test a simple codon usage program on your DNA sequence.

1. **Point your browser to** `bioweb.pasteur.fr`**.**

   Click the English/American flag (top right) for the English version.

2. **Click the EMBOSS link at the bottom-left of the page.**

   The list of EMBOSS modules appears.

3. **Click the wordcount link.**

   The wordcount page appears in all its glory.

4. **Enter your e-mail address and paste your sequence (raw format is okay) in the appropriate fields. Be sure to enter the word size you're interested in as well.**

5. **Click the Run wordcount button at the top-left of the page.**

   After a few seconds, a result page appears.

6. **Click the wordcount.out link to get your results as a simple list.**

Experimenting with other DNA composition analyses

While you're at the Pasteur Institute site, use the opportunity to further explore the list of available EMBOSS modules, such as *chips* (for codon usage statistics), or the CpG rich region finder *cpgreport.*

Finding internal repeats in your sequence

Another useful type of composition analysis involves locating segments that occur more than once within your sequence. Such segments are called *repeats.* There is no real difference between long words (6-tuple, 8-tuple, and larger) and repeats. Here's a common-sense rule: *A repeat is a word long enough so that it's unlikely to occur very often by chance, given a random sequence.* For instance, a GTC triplet found 4 times within a 500-nucleotide long sequence doesn't qualify as a repeat.

Another difference between word-counting and repeat analysis is that repeats can be imperfect. Unlike words, similar repeats don't need to be identical. Finally, biologists like to distinguish *tandem repeats* (similar subsequences along the same DNA strand) from *inverted repeats* (similar sequences occurring on the direct and reverse strands). Biologists are interested in repeats because they are often involved in genome rearrangements or regulatory mechanisms of gene expression.

There are many different algorithms for finding repeats within a DNA (or protein) sequence. They all try to identify segments more similar to each other than would be expected by mere chance alone. The tricky part is in the scoring and ranking of the similar subsequence segments. Is the exact matching of five consecutive nucleotides good enough to be considered a repeat? And is it better than 9 out of 10, or 123 over 160? Which one do you want reported first? How far down the list of possible repeats do you want to go?

Finding repeats is a tricky business

Because there are no universal answers to the questions surrounding the precise nature of repeats, repeat-finding programs ask you to fix thresholds related to their scoring algorithms, repeat size, copy number, periodicity (distance between repeats), and other things you don't always understand. This makes them difficult to use. The default settings they provide may or may not work for your particular sequence.

In that respect, our quick survey of Web-based repeat finders — using a repeat-containing sequence we made just for this purpose — was amazingly unsuccessful on all the available servers. But that doesn't mean these servers are bad — only that the problem is truly complicated.

Never believe a negative output from a repeat finder!

Fortunately, there is one simple universal (DNA/protein) approach that avoids all these difficulties. It is the *dot-plot* approach, which we show you in the following steps list. (For more on dot plots, see Chapter 8.)

For long nucleic acid sequences, you can use the dot-plot program provided by the Molecular Toolkit Web service at Colorado State University. Just do the following:

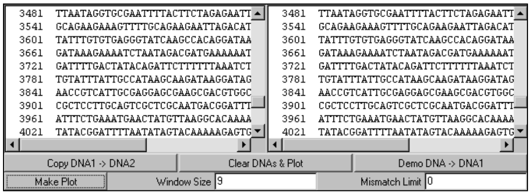
1. **Point your browser to** `arbl.cvmbs.colostate.edu/molkit/`**.**

   The Molecular Toolkit is a group of programs for analysis and manipulation of nucleic acid and protein sequence data. The programs are written in Java (1.0), and require that your browser support this language.

   This site is very useful if you need to perform some simple DNA sequence manipulation, such as reverse complementation (changing strands), protein translation, or if you want to display a restriction map. For now, continue through these steps.

2. **Click the** Dot Plots **link at the top of the list.**

   You obtain a straightforward input form, with two side-by-side boxes titled DNA Number 1 and DNA Number 2, as shown in Figure 5-5.



```
3481  TTAATAGGTGCGAATTTTACTTCTAGAGAATT
3541  GCAGAAGAAAGTTTTGCAGAAGAATTAGACAT
3601  TATTTGTGTGAGGGTATCAAGCCACAGGATAA
3661  GATAAAGAAAATCTAATAGACGATGAAAAAAT
3721  GATTTTGACTATACAGATTCTTTTTTAAATCT
3781  TGTATTTATTGCCATAAGCAAGATAAGGATAG
3841  AACCGTCATTGCGAGGAGCGAAGCGACGTGGC
3901  CGCTCCTTGCAGTCGCTCGCAATGACGGATTT
3961  ATTTCTGAAATGAACTATGTTAAGGCACAAAA
4021  TATACGGATTTTAATATAGTACAAAAAGAGTG
```

```
3481  TTAATAGGTGCGAATTTTACTTCTAGAGAATT
3541  GCAGAAGAAAGTTTTGCAGAAGAATTAGACAT
3601  TATTTGTGTGAGGGTATCAAGCCACAGGATAA
3661  GATAAAGAAAATCTAATAGACGATGAAAAAAT
3721  GATTTTGACTATACAGATTCTTTTTTAAATCT
3781  TGTATTTATTGCCATAAGCAAGATAAGGATAG
3841  AACCGTCATTGCGAGGAGCGAAGCGACGTGGC
3901  CGCTCCTTGCAGTCGCTCGCAATGACGGATTT
3961  ATTTCTGAAATGAACTATGTTAAGGCACAAAA
4021  TATACGGATTTTAATATAGTACAAAAAGAGTG
```

Window Size 9

Mismatch Limit 0

Figure 5-5: Input form of the DNA dot-plot program.

3. **Copy your sequence from a .txt file or a Word document, by using Ctrl+C or the Copy button on the Word toolbar.**

4. **Paste your sequence in the DNA Number 1 window by using Ctrl+V or Word's Paste button.**

   The program is happy with the raw sequence format (nucleotide only).

5. **Copy your sequence into the DNA Number 2 window as well.**

   For this, you can use the special Copy DNA1 -> DNA2 button provided below the DNA Number 1 window.

6. **Click the Make Plot button, located below the DNA sequence boxes.**

   Your dot plot magically appears. That's all there is to it!

   Limit yourself to a sequence under 10,000 bp if you aren't prepared to wait for several minutes.

In this case, you can get the dot plot quickly — in real time. It appears in the square below the input windows. For the example in Figure 5-6, we made up a sequence in which a large segment is imperfectly repeated three times. These repeats show up as three lines both in the upper and lower diagonal parts of the plot. The locations of the repeated sequence segments are found by projecting them onto the main diagonal (see Figure 5-6). The long repeats stand out in comparison to non-significant, shorter, random repeats of length 9 to 15 (which show up as mere dots). Using your own sequence, you can experiment with the Window Size parameters (minimal size of the repeat) as well as the Mismatch Limit parameters (number of non-identical nucleotides) to see how they influence the graph appearance.

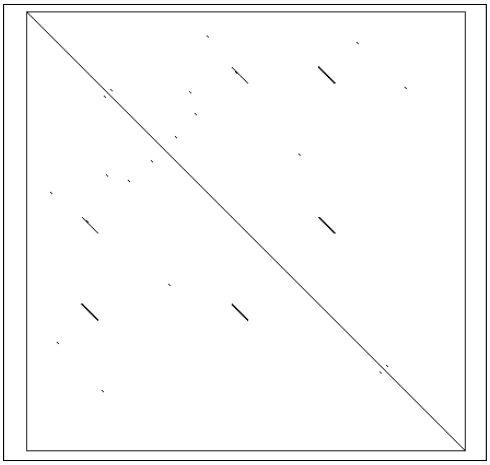


Figure 5-6: Dot plot showing a segment repeated three times.

Using dot plots to identify inverted repeats

It's important to remember that you can also use a dot-plot program to search for inverted repeats — local similarities between the *direct* and *reverse* strand of a DNA sequence. To do this, you simply have to use the sequence along the X-axis (DNA Number 1), and its *reverse-complement* along the Y-axis (DNA Number 2). Use the Manipulate Sequences option at `arbl.cvmbs.colostate.edu/molkit/` to generate a sequence's *reverse-complement.* The only difference with the previous protocol is that the main diagonal disappears. Inverted repeats will show up again as off-diagonal segments.

Identifying genome-specific repeats in your sequence

WARNING!

Do not confuse the discovery of a new repeat in your sequence with the identification of a repeat from a pre-established list of repeats, like the Alu repeat family in the human genome. Discovering a new repeat has to do with the internal structure of your sequence; identifying an established one is simply a matter of recognizing some local similarities between your sequence and a predefined reference database of repeats.

Repeated elements mostly occur in multicellular organisms. Vertebrate genomes are particularly rich in many families of repeated elements of various sizes. You can get a fairly complete picture of this topic by visiting the Repbase Web site of the Genetic Information Research Institute at `www.girinst.org`. This site gives you access to the CENSOR software tool, which screens query sequences against a reference collection of repeats, "censors" (that is, masks) matching portions with special symbols, and also generates a report on found repeats.

A similar service is offered at `www.repeatmasker.org` on a server based at the Institute for Systems Biology. Note that the program will mask, on average, a whopping 50% of the human genome sequence!

Finding Protein-Coding Regions

Get ready — this is where the fun starts! We've checked our DNA sequence for contamination, verified its restriction maps, computed various composition statistics, and identified its repeats. It's time to move to the really exciting stuff! We can search to find whether and where a protein is encoded in that obscure ATGCTACG gobbledygook.

To understand what is known as *protein discovery,* you must remember that protein-coding genes have vastly different structures in microbes and multicellular organisms. In microbes, each protein is encoded by a simple DNA segment — from start to end — called an open reading frame (ORF). In animals and plant genes (vertebrates are the worst), proteins are encoded in several pieces called *exons,* separated by non-coding DNA segments called *introns.* That explains why the methods and programs used for finding proteins in microbes and higher eukaryotes are rather different.

This section starts with a simple strategy called *ORFing,* which shows you how to find protein-coding regions in microbial DNA sequences or eukaryotic mRNA sequences.

ORFing your DNA sequence

To code for a protein, a DNA sequence must contain a translation Start codon (usually ATG) and not exhibit any of the Stop codons (usually TAA, TAG, TGA) in phase with the ATG for quite some length. This is the precise definition of an open reading frame. Given that proteins have an average length of about 350 residues (and that very few of them are smaller than 100 residues), you can use an additional criterion — minimal size — for example, requesting that at least 300 nucleotides separate the Start from the Stop. This is what ORFing is all about. You can get some practice by using ORF Finder at NCBI (the National Center for Biotechnology Information), as we describe in the following steps list:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/gorf/gorf.html`**.**

   The Input form is made up of a simple box, where you must paste your sequence (raw format), and a pull-down menu of genetic code choices. (We're ready to bet good money that you never knew there were 16 genetic code choices. So much for the universal code!) For now, let's stick with the standard genetic code (the default option).

2. **Copy your sequence from a .txt file or a Word document.**

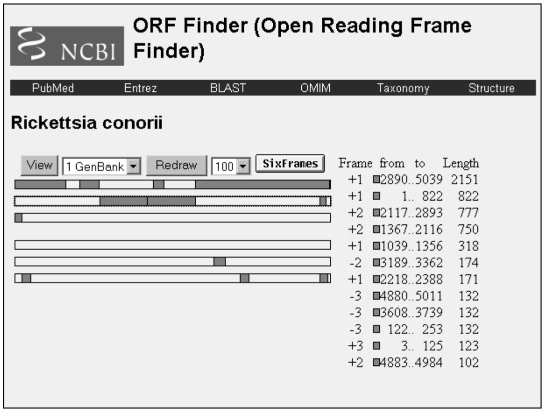
   We used the first 5,000 bp of GenBank entry AE008569 (*Rickettsia conorii* genome). Alternatively, you can enter this accession number in the relevant input box, and indicate from 1 to 5,000 below the sequence input box. Then directly go to Step 4.

3. **Paste your sequence in the input box.**

   The program is happy with the raw sequence format (nucleotide only).

4. **Click the OrfFind button.**

   Figure 5-7 shows a typical output.



Rickettsia conorii

```
Frame  from    to   Length
 +1   2890..5039  2151
 +1      1..  822   822
 +2   2117..2893   777
 +2   1367..2116   750
 +1   1039..1356   318
 -2   3189..3362   174
 +1   2218..2388   171
 -3   4880..5011   132
 -3   3608..3739   132
 -3    122..  253   132
 +3      3..  125   123
 +2   4883..4984   102
```

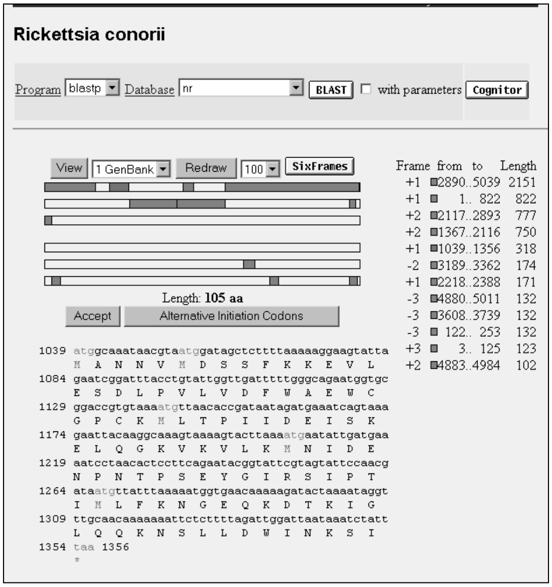
Figure 5-7: Typical output of the ORF Finder program.

Your DNA sequence is displayed as six parallel horizontal bars, with each one corresponding to one of the six possible translation frames: +1, +2, +3 and on the reverse strand: –1, –2, and –3. The ORFs verifying the default size threshold (100 nucleotides) appear as small green rectangles on each bar. In addition to the graphical representation, the ORFs are also listed according to size. At this point, you're free to modify the size threshold (50, 100, 300).

Always click the Redraw button to refresh the output.

To examine each of the ORFs more closely, click the corresponding rectangle in the graphical display or in the list to the side. This expands the output to include the predicted amino-acid sequence and some supplementary buttons, as shown in Figure 5-8. These buttons let you screen some of the NCBI databases for sequences similar to this ORF, *right on the spot!* This is a great, simple tool for ORFing your sequence.

Before getting into more complicated programs, we want to remind you that this simple ORFing program is also good for finding protein-coding regions for higher organisms *if your sequence is a cDNA.* cDNAs (the image of mRNAs) don't include introns — and they have a simple, microbe-like ORF structure. So you don't need to use another, more sophisticated program if you only want to delineate the protein-coding region within a human cDNA/mRNA sequence.



Rickettsia conorii

Program blastp Database nr ☐ with parameters

1 GenBank 100

Length: **105 aa**

Frame	from	to	Length
+1	2890	5039	2151
+1	1	822	822
+2	2117	2893	777
+2	1367	2116	750
+1	1039	1356	318
-2	3189	3362	174
+1	2218	2388	171
-3	4880	5011	132
-3	3608	3739	132
-3	122	253	132
+3	3	125	123
+2	4883	4984	102

```
1039 atggcaaataacgtaatggatagctcttttaaaaaggaagtatta
     M  A  N  N  V  M  D  S  S  F  K  K  E  V  L
1084 gaatcggatttacctgtattggttgattttgggcagaatggtgc
     E  S  D  L  P  V  L  V  D  F  W  A  E  W  C
1129 ggaccgtgtaaaatgttaacaccgataatagatgaaatcagtaaa
     G  P  C  K  M  L  T  P  I  I  D  E  I  S  K
1174 gaattacaaggcaaagtaaaagtacttaaaatgaatattgatgaa
     E  L  Q  G  K  V  K  V  L  K  M  N  I  D  E
1219 aatcctaacactccttcagaatacggtattcgtagtattccaacg
     N  P  N  T  P  S  E  Y  G  I  R  S  I  P  T
1264 ataatgttatttaaaaatggtgaacaaaaagatactaaaataggt
     I  M  L  F  K  N  G  E  Q  K  D  T  K  I  G
1309 ttgcaacaaaaaaattctcttttagattggattaataaatctatt
     L  Q  Q  K  N  S  L  L  D  W  I  N  K  S  I
1354 taa 1356
     *
```

Figure 5-8: Secondary output of the NCBI ORF Finder.

Analyzing your DNA sequence with GeneMark

The simplest ORFing protocols can probably correctly identify 85 percent of the protein-coding regions you may be interested in. However, there are a variety of situations that frequently occur where you may need to use a more sophisticated approach — the approach taken by GeneMark, for example. Such situations include

- Finding very short proteins
- Resolving ambiguous cases where overlapping ORFs are predicted in different reading frames — on the direct and reverse strand, for instance
- Wanting to pinpoint the exact Start codon (the most distal ATG isn't always the correct one)

GeneMark searches for coding regions using a criterion that's a bit more sophisticated than "it has to be an uninterrupted reading frame longer than a certain length." This program also takes into account the statistical properties (very similar to word/codon usage) of your sequence and associates some sort of a probabilistic quality index to each candidate's ORFs. In the process, some small ORFs may get promoted, and the precise Start location may be redefined. The quality index measures the similarity between the candidate ORFs and an ideal gene model. This gene model is often implemented as a Markov model, a mathematical concept beyond the scope of this book. The good news for you here is that these programs are easier to use than to understand. This is what we show you in the upcoming steps. Here goes:

1. **Point your browser to** `exon.gatech.edu/GeneMark/`**.**

   This is the main GeneMark home page, coming to us from Georgia Tech in Atlanta. It offers different specialized versions of the program (each corresponds to a different gene model) for working on prokaryotic (microbes), eukaryotic (animals), cDNA (the DNA version of mRNA), or virus sequences. (For the purpose of the example, let's say you want to analyze a sequence from a microbe that is little known.)

2. **In the Bacteria/Archaea section, click the** Heuristic models **link.**

   This selects the program version corresponding to the analysis of sequences from a new organism. It brings you to a simple form with the usual sequence input box.

3. **Copy your sequence from a** `.txt` **file or a Word document.**

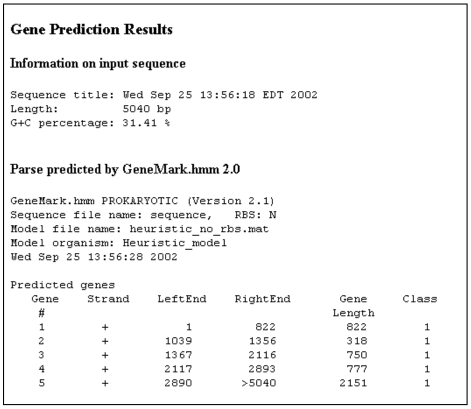
   We recommend using a microbial sequence about 5,000 bp in length.

4. **Paste the sequence in the input box.**

   The program is happy with the raw sequence format (nucleotides only) as well as many others.

5. **Click the Start GeneMark.hmm button to run the search.**

   After a while, the program outputs a list of predicted genes — with their locations — in a very simple format, as shown in Figure 5-9.



Gene Prediction Results

Information on input sequence

```
Sequence title: Wed Sep 25 13:56:18 EDT 2002
Length:         5040 bp
G+C percentage: 31.41 %
```

Parse predicted by GeneMark.hmm 2.0

```
GeneMark.hmm PROKARYOTIC (Version 2.1)
Sequence file name: sequence,   RBS: N
Model file name: heuristic_no_rbs.mat
Model organism: Heuristic_model
Wed Sep 25 13:56:28 2002

Predicted genes
   Gene    Strand    LeftEnd    RightEnd      Gene     Class
    #                                        Length
    1        +           1         822         822         1
    2        +        1039        1356         318         1
    3        +        1367        2116         750         1
    4        +        2117        2893         777         1
    5        +        2890       >5040        2151         1
```

Figure 5-9: Output of GeneMark.

If you're wondering what difference it makes to use a Markov model rather than a simple ORFing, the answer is in the differences between Figure 5-8 and Figure 5-9 — both of which relate to the same query sequence. GeneMark only retained the top five genes of the ORF Finder list!

Note that you can also run GeneMark (in the known model version) from its site at the European Bioinformatics Institute. The URL is

```
www.ebi.ac.uk/genemark/
```

Finding internal exons in vertebrate genomic sequences

When it comes to predicting proteins from DNA, the most challenging problem around is analyzing a piece of human genomic DNA sequence. We already mentioned that vertebrate genes have a mosaic structure, where small bits of the protein are encoded by exons separated by non-coding introns. If you're looking at a human genomic sequence, your first question should be: *Do I have a protein-coding exon somewhere in there?*

According to what molecular biologists have worked out, a protein coding exon is an ORF flanked by two specific signals known as *splice sites.* Several programs exist that are meant to recognize these exons. They all work on the same principle: On a first pass, they identify candidate ORFs with good compositional properties (such as word frequency and codon usage). This is basically what GeneMark does. On a second pass, they retain only those that are exactly flanked by good splice-site sequences.

Vertebrate exons are small (150-bp long on average), and the sequences of their splice sites are variable. For this reason, finding exons is a more challenging problem than ORFing microbe DNA — so don't expect an exon-detection program to work nearly as well as something like ORF Finder or GeneMark. (After all, those programs have much larger targets to shoot at.) Still, we'd like to show you what an exon detection program can do. Check out the following steps, where we use Michael Zhang's program MZEF:

1. **Point your browser to `rulai.cshl.edu/`.**

   This is the Zhang Laboratory home page at Cold Spring Harbor Laboratory, on beautiful Long Island.

2. **On the next page, click the Gene-Finding link in the Software Tools section.**

3. **In the next Gene Finder page, select Human.**

   This selects the program version calibrated for human coding-region statistics. A simple input form is then displayed.

4. **Copy your sequence from a .txt file or a Word document.**

   If you don't have a sequence handy, you can fetch the sequence AF018429 from GenBank at NCBI (`www.ncbi.nlm.nih.gov`). This entry contains the exons 1 and 2 of the dUTPase gene.

   Make sure you get the sequence in FASTA format.

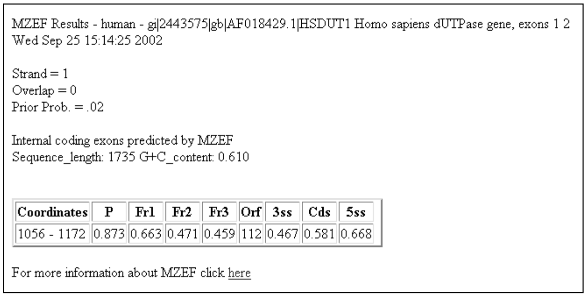
5. **Paste the sequence you copied into the input box.**

6. **Click the Submit button (below the sequence-input box) to start your analysis.**

   The program quickly returns a minimally formatted output, as shown in Figure 5-10. MZEF correctly identified one exon between positions 1056–1172. The rest of the line consists of various quality values associated with the overall prediction, the various reading frames (FR1, FR2, FR3), the splice site, and the coding potential. The prediction is half correct because it missed one exon. Furthermore, the predicted exon actually starts at position 1018. Such a success rate — about half the attempts, which isn't entirely perfect — isn't so unusual for exon-finding programs.

An exon predictor that combines MZEF with other approaches (which enhances its performances) is available at this Michigan Tech site:

```
genome.cs.mtu.edu/aat/aat.html
```



MZEF Results - human - gi|2443575|gb|AF018429.1|HSDUT1 Homo sapiens dUTPase gene, exons 1 2
Wed Sep 25 15:14:25 2002

Strand = 1
Overlap = 0
Prior Prob. = .02

Internal coding exons predicted by MZEF
Sequence_length: 1735 G+C_content: 0.610

Coordinates	P	Fr1	Fr2	Fr3	Orf	3ss	Cds	5ss
1056 - 1172	0.873	0.663	0.471	0.459	112	0.467	0.581	0.668

For more information about MZEF click here

Figure 5-10: MZEF output on the GenBank AF018429 sequence entry.

Complete gene parsing for eukaryotic genomes

If you ever get into serious genome sequencing, you'll need the highest level of sophistication in computer-assisted annotation: genome-parsing software. These programs are designed to take large (100,000 to several million bp) pieces of a genomic sequence at once and predict the detailed exon/intron structure of entire genes.

Like the MZEF program, these programs have a modular structure, where each module has been designed to recognize a given gene component: coding exons, first/last exons, promoter regions, poly-adenylation sites, and so on. The results of these independent modules are then combined into coherent gene-structure predictions (limiting exon splicing to compatible reading frames, for instance) — and these are finally scored according to their similarity with an ideal gene model. Markov models and dynamic programming optimization are what make these programs so effective.

The most recent genome parsers, such as GenomeScan, also take into account information on protein-sequence similarity.

Despite their increasing algorithmic complexity, these programs remain easy to use; all you really have to do is paste your sequence into an input window, click a button, and voilà — you have parsed genes!

Analyzing your sequence with GenomeScan

To show you how simple it is, we're going to show you how to use the GenomeScan parsing software program. To get things ready, prepare your own vertebrate genome sequence in FASTA format. It has to be long enough (100,000 bp) to contain at least one complete gene. Alternatively, you can use the demonstration sequence available on the GenomeScan site.

You also need protein sequences (also in FASTA format) that you know have some significant similarity with potential coding regions in your DNA sequence. You can find such proteins by doing a blastx comparison of your sequence to all known proteins. (For more on blastx, see Chapter 7.) Alternatively, you can use the demonstration set available on the GenomeScan site.

1. **Point your browser to** `genes.mit.edu/genomescan/`**.**

   The GenomeScan home page at the Massachusetts Institute of Technology duly appears. GenomeScan is the successor to Chris Burge's GenScan. The main difference is the use of protein homology information.

2. **Click the** GenomeScan Webserver **link to call up the page with the input form.**

3. **Choose Vertebrate from the Organism pull-down menu.**

   This selects the program version calibrated for human coding regions statistics.

4. **Copy your DNA sequence from a** `.txt` **file or a Word document.**

   Don't forget the FASTA header.

5. **Paste the DNA sequence you copied into the DNA Sequence input box — the upper of the two larger input boxes.**

   Alternatively, you may click the DNA testfile link at the top of the page to use the demo sequence.

6. **Copy your homologous protein sequence set from a** `.txt` **file or a Word document.**

   Alternatively you may click the protein file link at the top of the page to use the demo protein sequences.

   REMEMBER

   Each sequence must have a FASTA header. The set may contain multiple proteins so long as each is separated by a header on its own line. Files should contain less than 1 million bases.

7. **Paste your protein sequence into the Protein Sequence input box — the lower of the two larger input boxes.**

8. **Click the Run GenomeScan button at the bottom of the form to start the analysis.**

The output is a very long table listing all the components of the various predicted genes, with their coordinates and associated quality indices. This output is meant to be read by computer programs in the context of large-scale sequencing projects and automated annotation. Fortunately, GenomeScan also provides a nice graphical output both as PDF and PostScript images.

Figure 5-11 shows an example of the graphical output generated from the demo DNA and protein sequences (sorry, you'll have to use your imagination to provide the colors on this black-and-white page). Predicted genes (red arrows) and their predicted exons (red rectangles) are clearly indicated, together with their supportive evidence from blastx similarity (green rectangle below). A total of 5 genes are predicted along this 100,000-bp long vertebrate genomic sequence.

Assembling Sequence Fragments

Depending on their technology, current sequencing machines can only produce sequences at lengths of 500 to a 1,000 nucleotides at a time. Thus, if your DNA sequence is longer than this, it has been obtained by assembling shorter, overlapping fragments. Research into the various assembly algorithms — as well as research devoted to the development of better assembly software — is vital to bioinformatics.

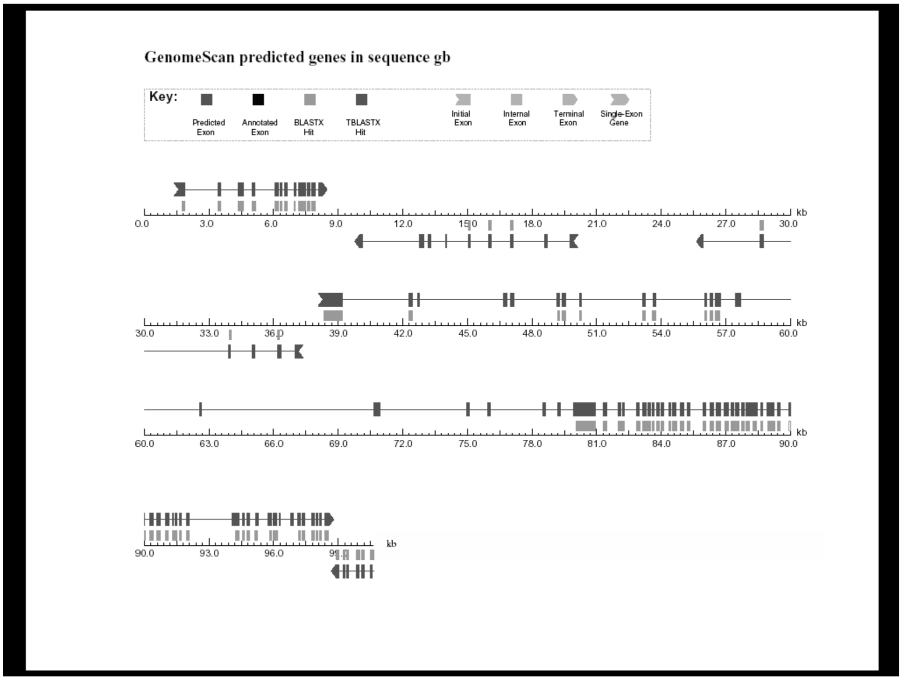

GenomeScan predicted genes in sequence gb
Key:
Predicted Exon
Annotated Exon
BLASTX Hit
TBLASTX Hit
Initial Exon
Internal Exon
Terminal Exon
Single-Exon Gene
0.0 3.0 6.0 9.0 12.0 15.0 18.0 21.0 24.0 27.0 30.0 kb
30.0 33.0 36.0 39.0 42.0 45.0 48.0 51.0 54.0 57.0 60.0 kb
60.0 63.0 66.0 69.0 72.0 75.0 78.0 81.0 84.0 87.0 90.0 kb
90.0 93.0 96.0 99.0 kb

Figure 5-11: Example of GenomeScan graphical output.

Managing large sequencing projects with public software

Taking what we would label the shotgun approach, most researchers stitch together a typical microbial genome sequence (4MB) using over 50,000 individual fragments known as *reads.* The programs used for such genome sequencing — or for bigger projects involving plant or animal genomes — take their input directly from the sequencing machine's *chromatograms* or *traces.* These traces are complex fluorescent profiles with peaks and valleys revealing the order and nature (A, T, G, C) of each nucleotide in the reads. Such programs include a base-calling system that associates a quality score to each nucleotide at each position of each read. They also include a data-management system, as well as interactive editing and display tools. Because of their complexity, these complete genome-sequencing software packages are not interactively available on Web sites; you have to install them locally on a dedicated powerful machine. Using them efficiently also requires some serious study. The most popular genome-sequencing packages that are publicly available (for academics) are these:

- **The Staden Package:** `staden.sourceforge.net/`
- **Phred/Phrap/Consed:** `www.phrap.org/`
- **The Acembly engine included in the popular AceDB suite:** `www.acedb.org/`
- **The Assembler from The Institute for Genome Research (TIGR):** `www.tigr.org/software/`

Commercial genome-sequencing software packages are also available from several companies:

- **Sequencher (from Gene Codes):** `www.genecodes.com/`
- **Lasergene (from DNASTAR):** `www.dnastar.com/`

We believe that showing you how to manage a large genome-sequencing project is beyond the scope of this introductory book. However, you may still encounter the need to assemble a handful of sequences in your everyday research life. For instance, you might be working on a cDNA that's just a few thousand bp in length, represented by a dozen PCR fragments. You may have generated a number of *expression sequence tags* (ESTs — that is, pass-sequenced partial cDNAs) and want to see whether you can deduce a complete cDNA from the tags. Alternatively, you may want to assemble ESTs you extracted from a database. In the latter case, you may not have the corresponding traces and base-call quality scores.

What you're looking for is a simple program able to recognize significant overlaps between your fragments and assemble them into a single sequence, known as a *contig.* Of course, the fragment sequences might have been obtained from different DNA strands, making the visual recognition of overlapping segments quite difficult. The following section shows you how to do this with the CAP3 sequence assembly program (go to `genome.cs.mtu.edu/` for credits and more information on this freely available code). We'll use CAP3 as implemented on a public server based in Lyon, France.

Assembling your sequences with CAP3

Prepare the set of sequences that you want to assemble in FASTA format as a `.txt` file or a Word document. Each individual sequence must be in FASTA format and have its own header. For the sake of simplicity, we're going to use the small demo dataset provided with the CAP3 documentation at `genome.cs.mtu.edu/cap/data/`. Simply open the `seq` file and copy the six FASTA-formatted sequences (R1 to R6) using the Select All/ Copy options from your browser. We are now ready to assemble them with CAP3 on the Lyon bioinformatics platform.

1. **Point your browser to** `pbil.univ-lyon1.fr/cap3.php`**.**

   The simplest possible form appears, made of one input sequence window above a Submit button and a Clear button. The Lyon server allows a total of 50 kb of nucleotide sequences to be processed, which is usually plenty for the purpose of assembling a cDNA.

2. **Paste the set of sequence into the Sequence input box.**

   We used the CAP3 demo sequences R1 to R6.

   WARNING!

   Don't forget the FASTA headers! (This is the only way for the program to distinguish a fragment sequence from the next.)

   TIP

   It doesn't matter whether you provide a mixture of sequences in different orientations. The program tries them in both orientations automatically.

3. **Click the SUBMIT button to run the assembly.**

   The server runs CAP3 with the default setting. A full implementation would allow you to play with the length threshold and level of similarity of the overlaps requested to assemble fragments, and to associate a file of sequence quality to the sequence data. (See the CAP3 documentation for details.)

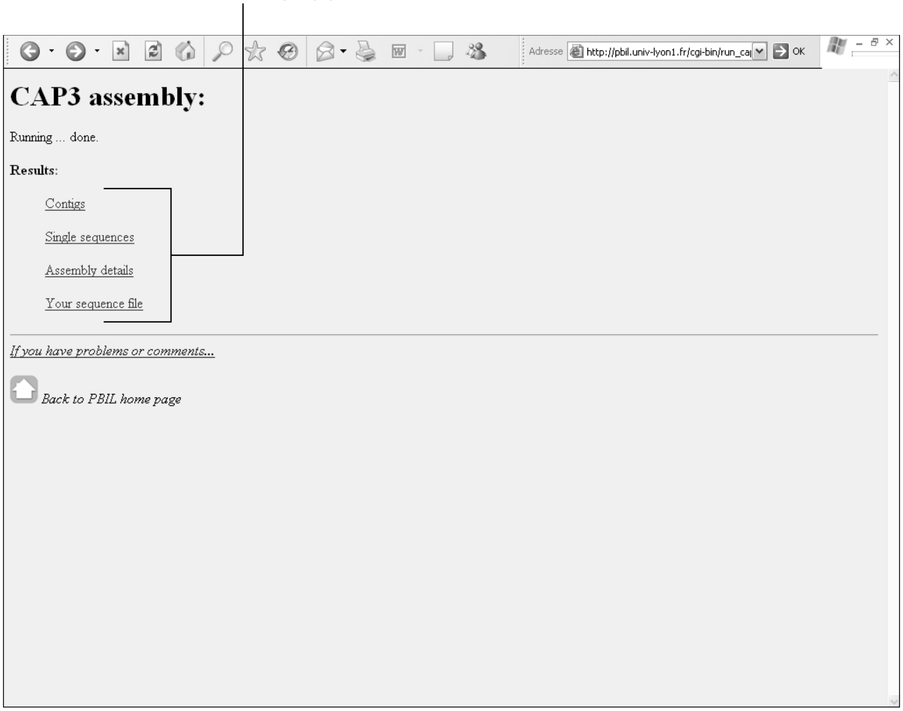
If you provided the demo sequences, the program response will be almost immediate, and the output will look like Figure 5-12, displaying four links:

- **Contigs:** Contains the final assembled sequence(s)
- **Single Sequences:** Contains the input sequence fragment not incorporated in the assembly.
- **Assembly Details:** The precise relationship between the fragments within the contigs.
- **Your Sequence File:** A summary of the input sequence fragments

Click each of these links in turn to see their content.

For instance, clicking the Assembly Details link displays the overlaps and contig structure. (See Figure 5-13.) One can see that R5 is not part of the assembly. It is in fact found in the Single Sequences file. The resulting contig consensus sequence is displayed by clicking the Contigs link. (Refer to Figure 5-12.)

Click on these links to display your results.



CAP3 assembly:

Running ... done.

Results:

Contigs

Single sequences

Assembly details

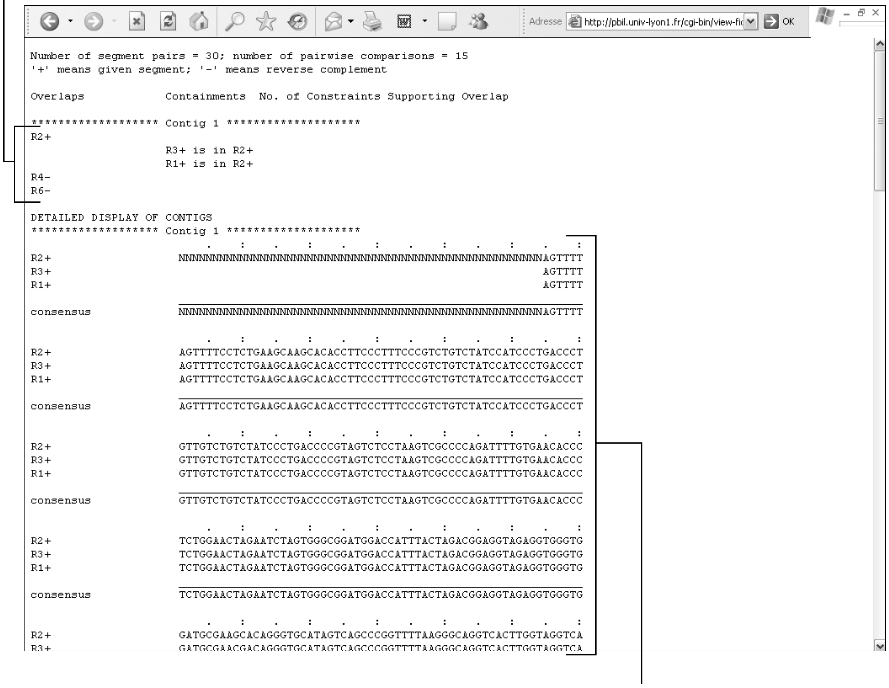
Your sequence file

If you have problems or comments...

Back to PBIL home page

Figure 5-12: Output files of the CAP3 assembly program.

List of assembled fragments in contig1



```
Number of segment pairs = 30; number of pairwise comparisons = 15
'+' means given segment; '-' means reverse complement

Overlaps            Containments  No. of Constraints Supporting Overlap

******************* Contig 1 ********************
R2+
                    R3+ is in R2+
                    R1+ is in R2+
R4-
R6-

DETAILED DISPLAY OF CONTIGS
******************* Contig 1 ********************
                          .    :    .    :    .    :    .    :    .    :    .    :
R2+                   NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGTTTT
R3+                                                                     AGTTTT
R1+                                                                     AGTTTT
                      ____________________________________________________________
consensus             NNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNNAGTTTT


                          .    :    .    :    .    :    .    :    .    :    .    :
R2+                   AGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCTGACCCT
R3+                   AGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCTGACCCT
R1+                   AGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCTGACCCT
                      ____________________________________________________________
consensus             AGTTTTCCTCTGAAGCAAGCACACCTTCCCTTTCCCGTCTGTCTATCCATCCCTGACCCT


                          .    :    .    :    .    :    .    :    .    :    .    :
R2+                   GTTGTCTGTCTATCCCTGACCCCGTAGTCTCCTAAGTCGCCCCAGATTTTGTGAACACCC
R3+                   GTTGTCTGTCTATCCCTGACCCCGTAGTCTCCTAAGTCGCCCCAGATTTTGTGAACACCC
R1+                   GTTGTCTGTCTATCCCTGACCCCGTAGTCTCCTAAGTCGCCCCAGATTTTGTGAACACCC
                      ____________________________________________________________
consensus             GTTGTCTGTCTATCCCTGACCCCGTAGTCTCCTAAGTCGCCCCAGATTTTGTGAACACCC


                          .    :    .    :    .    :    .    :    .    :    .    :
R2+                   TCTGGAACTAGAATCTAGTGGGCGGATGGACCATTTACTAGACGGAGGTAGAGGTGGGTG
R3+                   TCTGGAACTAGAATCTAGTGGGCGGATGGACCATTTACTAGACGGAGGTAGAGGTGGGTG
R1+                   TCTGGAACTAGAATCTAGTGGGCGGATGGACCATTTACTAGACGGAGGTAGAGGTGGGTG
                      ____________________________________________________________
consensus             TCTGGAACTAGAATCTAGTGGGCGGATGGACCATTTACTAGACGGAGGTAGAGGTGGGTG


                          .    :    .    :    .    :    .    :    .    :    .    :
R2+                   GATGCGAAGCACAGGGTGCATAGTCAGCCCGGTTTTAAGGGCAGGTCACTTGGTAGGTCA
R3+                   GATGCGAACGACAGGGTGCATAGTCAGCCCGGTTTTAAGGGCAGGTCACTTGGTAGGTCA
```

Contig building details and consensus sequence

Figure 5-13: CAP3 Assembly details (partial).

Beyond This Chapter

This chapter provides you with a good background on the various kinds of analyses people usually want to perform on their newly determined DNA sequences. There are, of course, many other things you can do and look for when you're working within a more specialized context. The most important area that isn't covered here concerns the search for biologically significant signals in the non-coding part of DNA sequences. This includes very hot topics — such as the prediction of promoter regions, regulatory elements, and/or protein binding sites.

Our main excuse for not presenting this area in more detail is that the success rate for these analyses is still quite low. Also, the best programs and relevant databases aren't always in the public domain or aren't available as Web servers. Table 5-1 provides you with some entry points you can explore for yourself.

Table 5-1 Free Web Sites for Searching Signals in DNA Sequences

Institution; Program	*URL*	*Description*
Zhang Lab, Cold Spring Harbor, USA	`rulai.cshl.edu/`	Search for promoters and various conserved elements
Center for Information Technology, NIH, USA; WWW Signal Scan	`bimas.dcrt.nih.gov/molbio/signal/`	Search for transcriptional elements
Center for Information Technology, NIH, USA; WWW Promoter Scan	`bimas.dcrt.nih.gov/molbio/proscan/`	Predict eukaryotic promoter regions
Munich Bioinformatics Center, GER, CREDO	`mips.gsf.de/services/analysis`	*Cis*-Regulatory Element Detection
San Diego Supercomputer Center, USA; MEME & MAST	`meme.sdsc.edu/meme/intro.html`	Discover motifs in groups of related DNA or protein sequences
Université Libre de Bruxelles, Belgium	`rsat.ulb.ac.be/rsat/`	Analyze regulatory sequences

Chapter 6

Working with a Single Protein Sequence

In This Chapter

- Knowing what you must about domains, HMMs, profiles, and the Pfam domain collection
- Visiting the three most popular sites for finding domains in your protein
- Predicting simple physical properties of your sequences
- Predicting protease digestion patterns
- Predicting coiled-coil domains
- Predicting post-translational modifications

> *Once you eliminate the impossible, whatever remains, no matter how improbable, must be the truth.*
>
> — Sherlock Holmes, as told to Sir Arthur Conan Doyle (1859–1930)

When you're studying a protein, you turn yourself into an investigator. Basically, you want to find out everything you can before designing a new experiment in the lab. For instance, in order to make sure that the protein you have in your test tube really is what you think it is, you must carry out a simple physical analysis. Many online tools exist that help you predict the results of these analyses (and make sure you're on the right track). In this chapter, we show you where you can find these tools and how to use them. Among other things, we show you how to predict molecular weights, measure isoelectric points, count residues, or predict the outcome of a protease digestion.

Of course, there's more to a protein than its weight or its charge. While these properties do help you get a sense of what's what with your particular protein, uncovering a nice biological story takes more than that. An important question to ask yourself is what your protein looks like when it's active. Does it need to be modified after being translated, does it contain coiled-coil elements, or is it a transmembrane protein? In this chapter, we show you how online tools can help you answer these questions.

The structure is also important, of course — so important that we've dedicated two chapters of this book to its analysis (Chapter 11 for protein structure analysis and Chapter 12 for RNA structure analysis). For this reason, we don't talk much about structures in this chapter.

Knowing what your protein looks like is one thing, but knowing what it actually does is another. For instance, you may want to know if your protein binds calcium or if it contains an enzymatic site. If it's an enzyme, there's no doubt that you need to know about its substrate (the kind of molecule it binds). To answer these types of questions, we show you three powerful methods to check whether your protein contains a domain with a known function.

The other powerful technique that can help you when it comes to guessing the function of a protein involves *similarity searches.* If, in a database somewhere, you find a well-characterized protein whose sequence is very similar to your protein, then you're allowed to say that most of what is true for this sequence is true for your sequence as well. Although we don't wade into this type of analysis in this chapter, don't worry — we won't leave you high and dry. We cover everything you need to know about similarity searches in Chapter 7.

Doing Biochemistry on a Computer

If you want to do some biochemistry using a computer — we're guessing you do — two terrific places to go are

- The ExPASy (*Ex*pert *P*rotein *A*nalysis *Sy*stem) server at `www.expasy.org`, with a specific page dedicated to protein analysis methods
- The Swiss EMBnet at `www.ch.embnet.org`

The people who created and maintain a big chunk of Swiss-Prot run these two sites. When it comes to looking closely at proteins, these folks know their jobs.

In this section, we assume that your heart is set on a protein you've never seen before. The only thing you know about this beauty is its sequence. Maybe this protein comes from a sequencing project or something similar. Before you open it up and check out the insides, you want to find out — and guess — as much about it as you can. That's the reason you're here reading this very informative chapter.

If you don't have your own protein yet, don't worry! For each example we use in this chapter, we provide you with a Swiss-Prot protein of your very own (the sequence, anyway). This way, you can always compare the results of the program with the Swiss-Prot annotation.

TIP

In general, when using a program for the first time, it's advisable to run it on a well-characterized protein, simply to check and see whether the program does what it says it does (and whether it does that well). To find a suitable example, use one of the SRS servers as we describe in Chapter 2. This can give you a sense of how good the program you're using really is.

Predicting the main physico-chemical properties of a protein

ProtParam, a program you can use online on the ExPASy server, is a convenient way to estimate every simple physico-chemical property (that is, each *prot*ein *param*eter) that can be deducted from your protein sequence. It makes no complex and adventurous assumptions about your little protein: It simply counts for you.

Here's how it works:

1. **Point your favorite browser to**

   ```
   www.expasy.org/tools/#primary
   ```

   The Primary Structure Analysis section of the ExPASy Proteomics Tools page appears, as shown in Figure 6-1.

   TIP

   At the end of this chapter, we give you an exhaustive list of all the available ExPASy mirror sites you can use if the main ExPASy site is down.

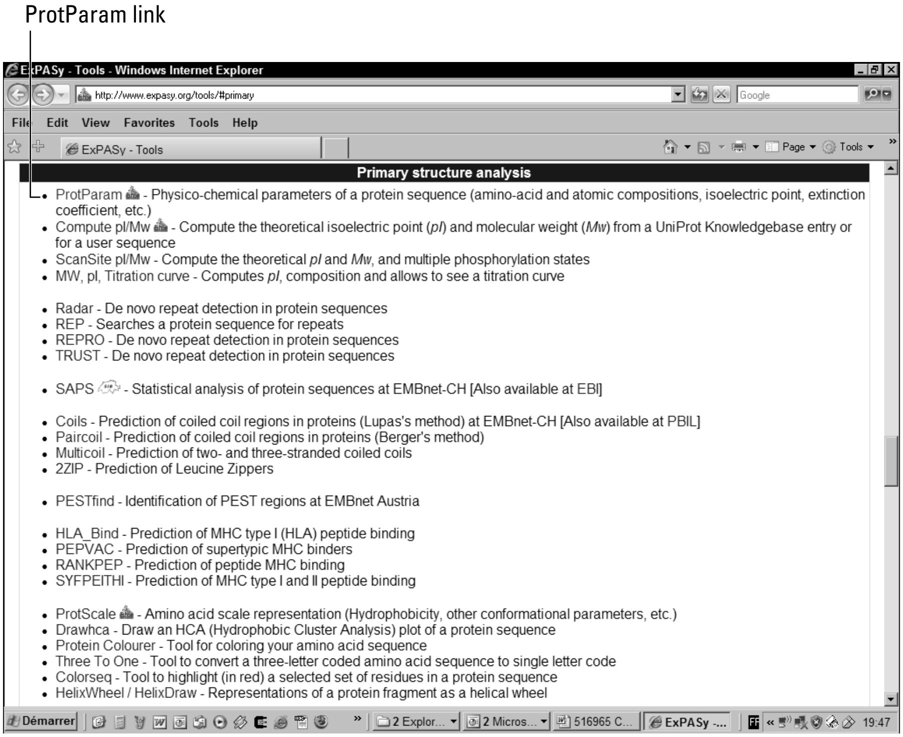
2. **Click the ProtParam link near the top of the page.**

   The ProtParam Tool page appears.

3. **Enter your sequence in the search boxes provided.**

   Provide a sequence in one of two ways:

   - Enter the accession number, as shown in Figure 6-2. You can do this if your sequence exists in the Swiss-Prot sequence database. Swiss-Prot accession numbers usually start with a letter followed by five digits. In this example, we use a Swiss-Prot protein: rat Syntaxin 1A (P32851).



ProtParam link

ExPASy - Tools - Windows Internet Explorer
http://www.expasy.org/tools/#primary
Google
File Edit View Favorites Tools Help
ExPASy - Tools
Page
Tools
Démarrer
2 Explor...
2 Micros...
516965 C...
ExPASy -...
19:47

Primary structure analysis

- ProtParam - Physico-chemical parameters of a protein sequence (amino-acid and atomic compositions, isoelectric point, extinction coefficient, etc.)
- Compute pI/Mw - Compute the theoretical isoelectric point (*pI*) and molecular weight (*Mw*) from a UniProt Knowledgebase entry or for a user sequence
- ScanSite pI/Mw - Compute the theoretical *pI* and *Mw*, and multiple phosphorylation states
- MW, pI, Titration curve - Computes *pI*, composition and allows to see a titration curve

- Radar - De novo repeat detection in protein sequences
- REP - Searches a protein sequence for repeats
- REPRO - De novo repeat detection in protein sequences
- TRUST - De novo repeat detection in protein sequences

- SAPS - Statistical analysis of protein sequences at EMBnet-CH [Also available at EBI]

- Coils - Prediction of coiled coil regions in proteins (Lupas's method) at EMBnet-CH [Also available at PBIL]
- Paircoil - Prediction of coiled coil regions in proteins (Berger's method)
- Multicoil - Prediction of two- and three-stranded coiled coils
- 2ZIP - Prediction of Leucine Zippers

- PESTfind - Identification of PEST regions at EMBnet Austria

- HLA_Bind - Prediction of MHC type I (HLA) peptide binding
- PEPVAC - Prediction of supertypic MHC binders
- RANKPEP - Prediction of peptide MHC binding
- SYFPEITHI - Prediction of MHC type I and II peptide binding

- ProtScale - Amino acid scale representation (Hydrophobicity, other conformational parameters, etc.)
- Drawhca - Draw an HCA (Hydrophobic Cluster Analysis) plot of a protein sequence
- Protein Colourer - Tool for coloring your amino acid sequence
- Three To One - Tool to convert a three-letter coded amino acid sequence to single letter code
- Colorseq - Tool to highlight (in red) a selected set of residues in a protein sequence
- HelixWheel / HelixDraw - Representations of a protein fragment as a helical wheel

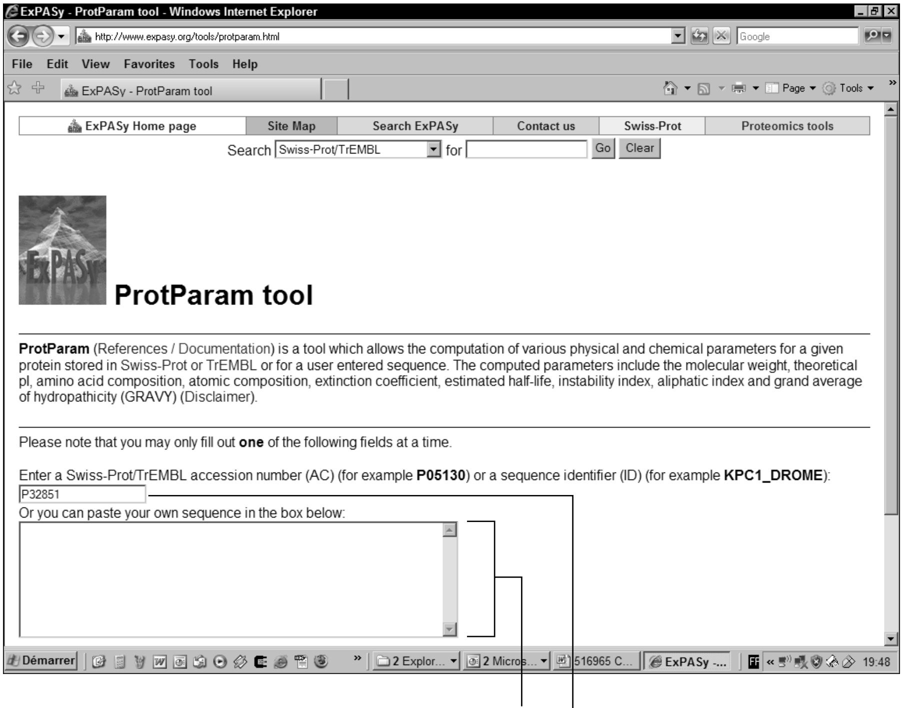
Figure 6-1: Working with the Primary Structure Analysis page.

- Paste the sequence — in raw format — into the window provided (see Figure 6-2 again). If your sequence isn't a Swiss-Prot sequence, or if you don't know the sequence name, you can paste the sequence itself, using this other window. Note that the sequence must be entered in *raw format* — it should not include anything but the standard abbreviations for amino acids (white space and numbers are tolerated but automatically removed). If your sequence is in FASTA format, DO NOT include the first line that starts with a greater-than sign (>).

4. **Click the Compute Parameters button.**

   If, in Step 2, you provided the accession number of the sequence, an intermediate page (like the one shown in Figure 6-3) appears.

5. **If you have entered a Swiss-Prot accession number — and Figure 6-3 has dutifully appeared — enter the range of the analysis in the N-Terminal and C-Terminal fields. (The N-terminus is the left side of your protein sequence and the C-terminus is the right side.)**



ProtParam tool

ProtParam (References / Documentation) is a tool which allows the computation of various physical and chemical parameters for a given protein stored in Swiss-Prot or TrEMBL or for a user entered sequence. The computed parameters include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index and grand average of hydropathicity (GRAVY) (Disclaimer).

Please note that you may only fill out **one** of the following fields at a time.

Enter a Swiss-Prot/TrEMBL accession number (AC) (for example **P05130**) or a sequence identifier (ID) (for example **KPC1_DROME**):
P32851

Or you can paste your own sequence in the box below:

Paste a raw sequence here.

Enter accession number here.

Figure 6-2: Entering an accession number into ProtParam.

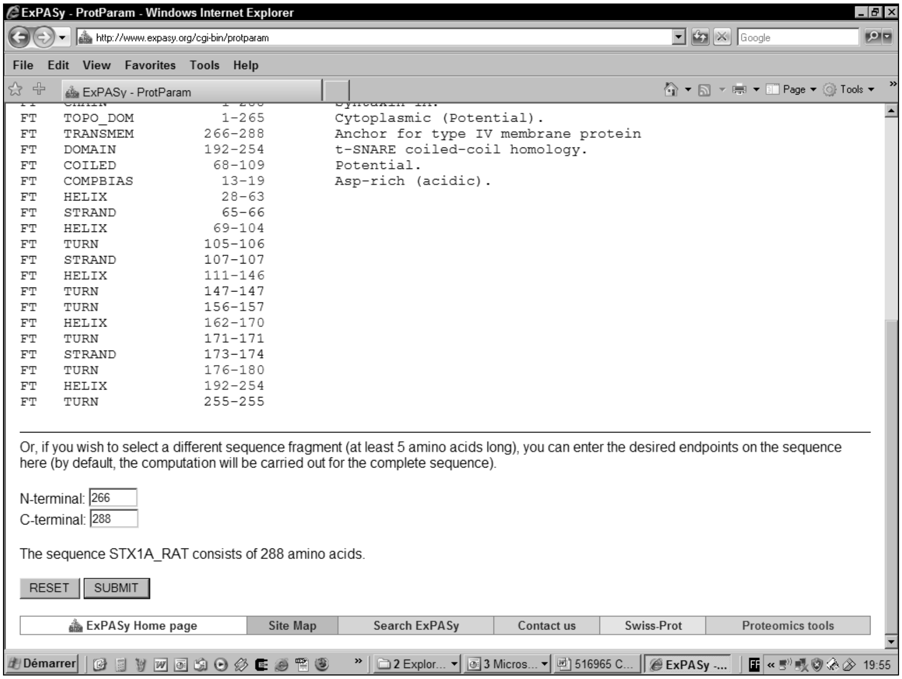
At the top of the page (partially) shown in Figure 6-3, you can find a list of the features contained in your sequence. This isn't a prediction — rather, it's a display of the information contained in the Swiss-Prot entry. Below are two boxes (N-terminal and C-Terminal) that you can fill with the coordinates of the segment you're interested in. If you leave these two boxes blank, ProtParam analyzes the complete sequence.

Using coordinates such as those in Figure 6-3, ProtParam analyzes the segment of your protein that starts on the 266th amino acid and ends on the 288th.

6. **Press the Submit button to proceed with the analysis.**

   The results — including molecular weight, extinction coefficient, and estimated half-life — appear on-screen.

7. **In your Web browser, choose File⇨Save As to save your results.**


ExPASy - ProtParam - Windows Internet Explorer
http://www.expasy.org/cgi-bin/protparam
Google
File Edit View Favorites Tools Help
ExPASy - ProtParam
Page
Tools

```
FT   CHAIN            1-288       Syntaxin 1A.
FT   TOPO_DOM         1-265       Cytoplasmic (Potential).
FT   TRANSMEM       266-288       Anchor for type IV membrane protein
FT   DOMAIN         192-254       t-SNARE coiled-coil homology.
FT   COILED          68-109       Potential.
FT   COMPBIAS        13-19        Asp-rich (acidic).
FT   HELIX           28-63
FT   STRAND          65-66
FT   HELIX           69-104
FT   TURN           105-106
FT   STRAND         107-107
FT   HELIX          111-146
FT   TURN           147-147
FT   TURN           156-157
FT   HELIX          162-170
FT   TURN           171-171
FT   STRAND         173-174
FT   TURN           176-180
FT   HELIX          192-254
FT   TURN           255-255
```

Or, if you wish to select a different sequence fragment (at least 5 amino acids long), you can enter the desired endpoints on the sequence here (by default, the computation will be carried out for the complete sequence).

N-terminal: 266
C-terminal: 288

The sequence STX1A_RAT consists of 288 amino acids.

Figure 6-3: Using ranges with ProtParam.

Interpreting ProtParam results

If you've used Syntaxin as your sample protein in your trial run of ProtParam, you can now proudly pass on the knowledge that Syntaxin's theoretical molecular weight is 33067,4 daltons, that its theoretical pI is 5.14, and that its aliphatic index is 83.96 — and this is only a sample of the large amount of information ProtParam can provide!

You can find detailed explanations on how ProtParam computes these parameters at `www.expasy.org/tools/protpar-ref.html`. Among other issues, the following sections draw your attention to a few points you may be interested in.

Molecular weight

The computer program is simply summing the weight of all the residues in the sequence. Remember that the program doesn't consider the following:

- It does NOT consider post-translational modifications such as glycosylations or phosphorylations.

- It does NOT take into account complex maturations such as the removal of a lead peptide or the excision of an intein.
- ProtParam does NOT know whether your mature protein becomes a dimer (association of two proteins), a trimer (association of three proteins), or any other kind of multimer.

Of course, these problems can eventually add up and lead to experimental results significantly different from the theory. In a way, that's where the fun begins . . . so keep your eyes open!

Extinction coefficients

An *extinction coefficient* tells you how much light (visible and invisible) your protein absorbs at a certain wavelength. Estimates of these values are useful if you need to follow your protein with a spectrophotometer when you purify it.

WARNING!

When you use this result, remember that the ProtParam value is only an indication. ProtParam predicts the extinction coefficient by summing the contribution of every amino acid as if each were independent. This calculation ignores the fact that the behavior of amino acids can be dramatically altered depending on their immediate surroundings. That behavior can't be predicted, so this estimate isn't exactly the last word (so to speak) on your protein's extinction coefficient.

If you need the *exact* coefficient, you must measure it experimentally. On the other hand, for most proteins, the experimental coefficient is (fortunately) rarely very different from the theoretical one.

Instability

This parameter is only just a crude estimate of your protein's stability in a test tube. When the index is below 40, the protein is usually stable. Above 40, it may not be stable.

Half-life

The *half-life* is a crude prediction of the time it takes for half of the amount of your protein present in a cell to disappear completely after its synthesis in the cell. This prediction is given for three types of organisms. You can safely extrapolate it to similar organisms.

WARNING!

The half-life value given here is meaningless if the degradation of your protein is part of a regulatory process. For instance, if the cell specifically directs a *protease* (a protein that digests other proteins) at your protein, your protein may disappear much faster than ProtParam tells you.

Digesting a protein in a computer

Protease digestions — where you use an enzyme to cut your protein in specific ways — can be useful if you're only interested in carrying out experiments on a portion of your protein. They can also be useful if you want to

- Separate the domains in your protein
- Identify potential post-translational modification by mass spectrometry
- Remove a tag protein when you express a fusion protein
- Make sure that the protein you're cloning isn't sensitive to some endogenous proteases

Fortunately, these days you can snip models of protein sequences in a computer — and see what happens *before* you start any hands-on lab procedures. PeptideCutter is a very useful tool for this kind of purpose — and is available from the ExPASy Web site at `www.expasy.org/tools/#proteome`. Check it out; you'll find it easy to use. Paste in your sequence, choose an option or two, click a button, and a model of your protein is sliced, diced, and chopped to your specifications.

Doing Primary Structure Analysis

The *primary structure* is the amino acid sequence of your protein. When you analyze the primary structure, you ignore the potential interactions between amino acids. Primary structure analysis of your protein doesn't tell you about the potential interactions between the amino acids in your sequence; you predict those when you predict the secondary or tertiary structure of your protein. (See Chapter 11 for more on such predictions.)

You conduct a primary structure analysis to try to find segments in your protein that display a special composition. These segments can reveal some interesting properties of your protein, such as

- *Hydrophobic regions* that could be membrane-spanning segments in proteins that anchor themselves into a membrane
- *Coiled-coil regions* that indicate potential protein-protein interaction
- *Hydrophilic stretches* that could be looping out at the surface of the protein

The methods we show you in this section, though less powerful than state-of-the-art methods of predicting secondary structure (which we cover in Chapter 11), still have a lot going for them. For one, they're simple to use and simple to understand. When you use them, you can easily see what's going on — and easily avoid making mistakes.

Sliding windows

The "sliding windows" technique is the most ancient way of looking at sequences. The principle is very simple. What you need is a chemical property and a list of values associated with each of the 20 amino acids. This property can be any measurable physico-chemical parameter, such as size, polarity, hydrophobicity, or even the propensity of amino acids to be in a specific structural state. The values in this table are the amino acids' *scale values.* Many such tables exist that have been determined experimentally for almost any characteristic you can think of.

After you have this table, you choose a window size and slide it along your sequence. When the window is centered on an amino acid, the scale values associated with the amino acids it contains are summed up and averaged. The resulting value is associated with the central amino acid, and then the window is shifted by one amino acid. This process goes on to the end of the sequence. The following example illustrates how the window slides along the sequence. The letter <--X--> indicates the amino acid on which the window is centered.

```
Sequence
   AGVCFGTRESALPTFREDCYGHZPLI
   KJFDESAQZ
     <---A---> Window 1
      <---G---> Window 2
       <---V---> Window 3
```

When the sliding operation is finished, the values associated with every amino acid are plotted against the sequence. Biologists name this display a property profile (do not confuse it with a domain profile, which is a formulation of a multiple sequence alignment). If you're lucky, you may be able to identify transmembrane segments, loops, or coiled-coil regions by using sliding windows. Hydrophobicity is the most popular analysis because it's a good indicator of transmembrane segments or core regions within a protein.

Many of the methods we show you here rely on the *sliding windows* technique. (For more on these guys, check out the "Sliding windows" sidebar.) Predictions based on sliding windows aren't very sensitive or very precise, but they are very robust. If you see a strong signal when using a method based on sliding windows, chances are you're looking at something that's a genuine biological signal.

Ironically, the main advantage of the methods that use sliding windows is also one of their main shortcomings: They don't interpret the results for you; they provide only a raw signal.

TIP

Because you have to do the interpretation yourself, two simple rules apply when interpreting the results of an analysis based on sliding windows:

- Be very strict and consider only strong signals.
- Check the robustness of your signal. Good signals aren't shy. They don't go away simply because you increase or decrease the window size by one amino acid or replace a property table with another similar table.

Another major weakness of the sliding-window method is arbitrary window size: why 9 residues instead of, say, 13 or 21? There is no magic answer to this question, but when it comes to choosing your window size, the following simple rule applies:

The window size must be in the same range as the size of the feature you're looking for.

For instance, if you're looking for transmembrane domains that are normally 21 amino acids long, you want to use a window size that is close to this value — in practice, 19 is the best window value for finding transmembrane regions. On the other hand, if you're interested in the structural features of globular proteins, a range between 7 and 11 would be more appropriate.

Looking for transmembrane segments

Predicting that your protein has transmembrane segments tells you much more about its function than almost any other simple prediction you can do.

When you know your protein is a transmembrane protein, you know that you can't work on it with the same techniques you'd use if it were a globular protein. If you find one transmembrane segment at the N-terminus of your sequence, you can predict that the protein is secreted. On the other hand, many transmembrane domains in one protein may indicate a channel.

Because these predictions are so important, we show you two methods for testing for transmembrane segments: ProtScale (a very simple one) and TMHMM (one of the most complete). TMHMM is not part of the ExPASy server; it is a service offered to the community by the Technical University of Denmark.

- *ProtScale* uses a sliding-window technique and one of many amino-acid scale values. In this example, we use the hydrophobicity to identify the groups of hydrophobic segments that characterize transmembrane proteins. ProtScale doesn't predict anything for you; it returns a hydrophobicity profile and lets you do the interpretation.
- *TMHMM* is a state-of-the-art program that predicts transmembrane segments in your protein. TMHMM also tells you about the portions of your protein that are probably inside the cell and those that are probably outside.

Running Protscale

Let's start with Protscale:

1. **Point your browser to** `www.expasy.org/cgi-bin/protscale.pl`**.**

   The ProtScale page duly appears.

2. **Enter your sequence accession number in the small search box.**

   You can paste the sequence in raw format or enter a Swiss-Prot Accession Number. (For an example of raw format, see Figure 6-2.)

   Here we've entered `P78588`, the accession number for `FREL_CANAL` protein, a protein that contains seven transmembrane segments.

3. **Scroll down the same page and then select the radio button next to Hphob. / Kyte & Doolittle, as shown in Figure 6-4.**

   ProtScale gives you a large range of properties that you can choose and test on your protein. The one we've chosen here is the most appropriate for predicting transmembrane helices.

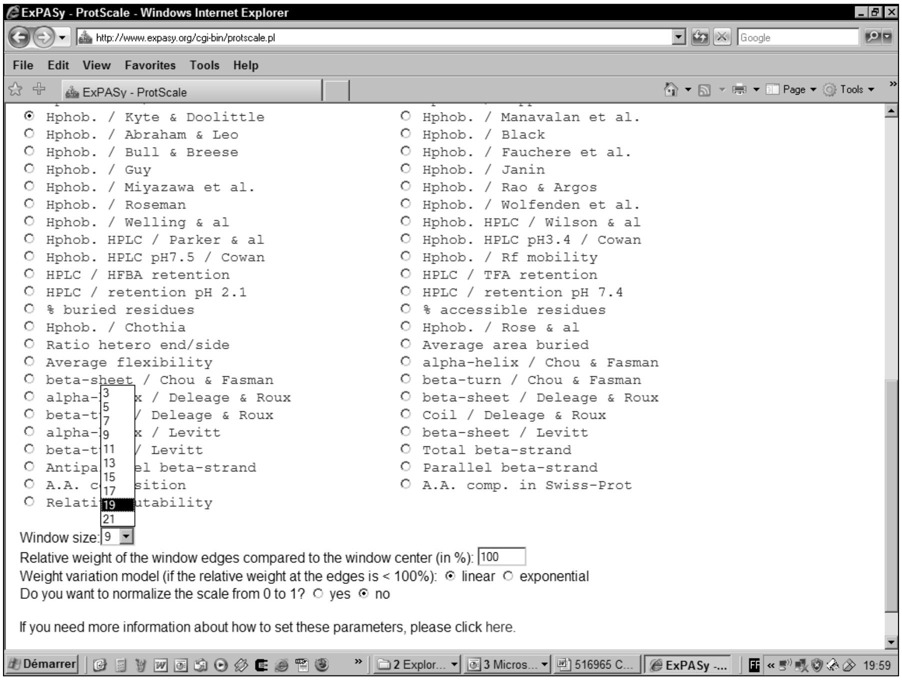
4. **From the Window Size pull-down menu, choose 19.**

   This value is appropriate when you're looking for transmembrane domains.

5. **Press the Submit button at the bottom of the page.**

6. **If you have entered a Swiss-Prot accession number, enter the range of the analysis.**

   If, in Step 2, you have provided the accession number of the sequence, an intermediate page appears like the one you saw in Figure 6-3.


ExPASy - ProtScale - Windows Internet Explorer
http://www.expasy.org/cgi-bin/protscale.pl
Google
File Edit View Favorites Tools Help
ExPASy - ProtScale
Page Tools
Démarrer
2 Explor...
3 Micros...
516965 C...
ExPASy -...
19:59

● Hphob. / Kyte & Doolittle
○ Hphob. / Abraham & Leo
○ Hphob. / Bull & Breese
○ Hphob. / Guy
○ Hphob. / Miyazawa et al.
○ Hphob. / Roseman
○ Hphob. / Welling & al
○ Hphob. HPLC / Parker & al
○ Hphob. HPLC pH7.5 / Cowan
○ HPLC / HFBA retention
○ HPLC / retention pH 2.1
○ % buried residues
○ Hphob. / Chothia
○ Ratio hetero end/side
○ Average flexibility
○ beta-sheet / Chou & Fasman
○ alpha- x / Deleage & Roux
○ beta-t / Deleage & Roux
○ alpha- x / Levitt
○ beta-t / Levitt
○ Antipa el beta-strand
○ A.A. c sition
○ Relati utability

○ Hphob. / Manavalan et al.
○ Hphob. / Black
○ Hphob. / Fauchere et al.
○ Hphob. / Janin
○ Hphob. / Rao & Argos
○ Hphob. / Wolfenden et al.
○ Hphob. HPLC / Wilson & al
○ Hphob. HPLC pH3.4 / Cowan
○ Hphob. / Rf mobility
○ HPLC / TFA retention
○ HPLC / retention pH 7.4
○ % accessible residues
○ Hphob. / Rose & al
○ Average area buried
○ alpha-helix / Chou & Fasman
○ beta-turn / Chou & Fasman
○ beta-sheet / Deleage & Roux
○ Coil / Deleage & Roux
○ beta-sheet / Levitt
○ Total beta-strand
○ Parallel beta-strand
○ A.A. comp. in Swiss-Prot

3
5
7
9
11
13
15
17
19
21

Window size: 9

Relative weight of the window edges compared to the window center (in %): 100

Weight variation model (if the relative weight at the edges is < 100%): ● linear ○ exponential

Do you want to normalize the scale from 0 to 1? ○ yes ● no

If you need more information about how to set these parameters, please click here.

Figure 6-4: Choosing a property on the ProtScale server.

At the top of the page in Figure 6-3, you can find a list of the features contained in your sequence. (This isn't a prediction, but rather a display of the information contained in the Swiss-Prot entry.) Below are two boxes that you can fill with the coordinates of the segment you're interested in. If you leave these two boxes blank, ProtParam analyzes the complete sequence.

7. **Click the Submit button to proceed with the analysis, and then wait for your browser to return the results.**

8. **When the Results page appears, click *Image in GIF format* at the bottom of the page.**

   A new page should appear on your browser, containing only the graph.

   Of the three formats proposed here, the GIF (*G*raphic *I*nterchange *F*ormat) is the most convenient if you want to include your graph in a presentation.

9. **Choose File⇨Save As from your browser's main menu to save your results.**

Interpreting ProtScale results

A good signal is strong and not very sensitive to the parameters. For instance, Figure 6-5 shows the (strong) results you obtain on the Swiss-Prot protein P78588 when using the Kyte and Doolittle Hydrophobicity Scale.

In order to convince yourself that this result is meaningful, redo the comparison with another hydrophobicity scale — like the one from Eisenberg, with the results you see in Figure 6-6. The details differ very little between these two graphs, but you can clearly see that the main features are well conserved and that the strongest peaks come at roughly the same positions.

TIP

In Figure 6-5, we've drawn a bold line that gives you a hint on how to interpret your results.

TIP

The recommended threshold value when using Kyte and Doolittle is 1.6 — but if you're like us and you keep forgetting this magic value, here is a simple recipe:

1. **Place a piece of paper over your results.**
2. **Lower this piece of paper until the tips of the strongest peaks appear.**
3. **Keep lowering this threshold as long as you can see nice sharp peaks.**

With this simple method, you can identify unambiguously five out of seven transmembrane regions. On a sixth one, we could make an adventurous guess — It's easy to guess like this when you know the answer! — but the seventh transmembrane domain is impossible to find.

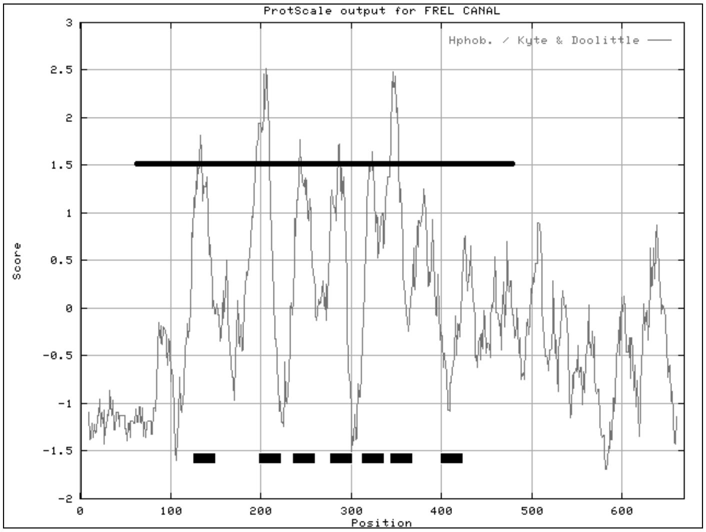


Figure 6-5: Hydrophobicity profile returned by ProtScale by using the Kyte & Doolittle Scale. The boxes mark known trans membrane segments.

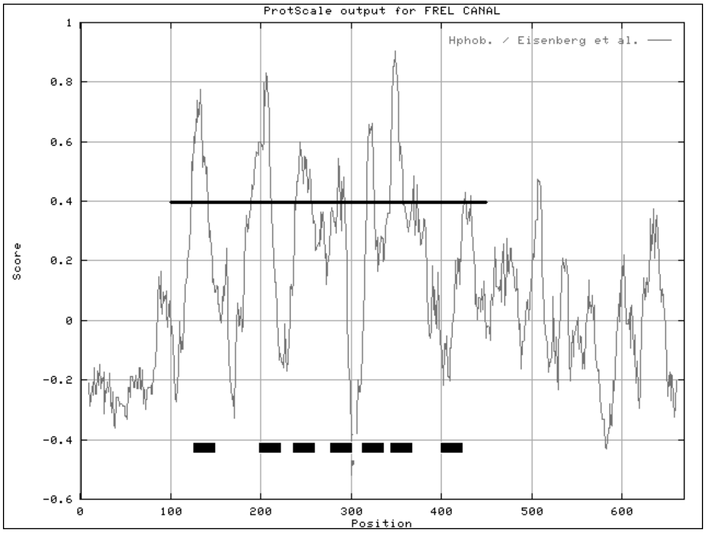


Figure 6-6: Hydrophobicity profile returned by ProtScale using the Eisenberg Scale.

That isn't so surprising: Many proteins contain these seven transmembrane regions — in which six are easy to find and the seventh domain is notoriously difficult to predict.

Running TMHMM

TMHMM is maintained by the Center for Biological Sequence Analysis (CBS) at the Technical University of Denmark. The CBS (www.cbs.dtu.dk/services/)

offers many interesting services for sequence analysis that we encourage you to explore.

TMHMM is a state-of-the-art method that uses sophisticated mathematical models (Hidden Markov Models) to predict transmembrane regions. For a bit of sophistication, give it a try:

1. **Point your browser to** `www.cbs.dtu.dk/services/TMHMM-2.0`**.**

   The TMHMM page of the CBS site appears (Figure 6-7).

2. **Scroll down the page a bit to enter your sequence in the TMHMM search box.**

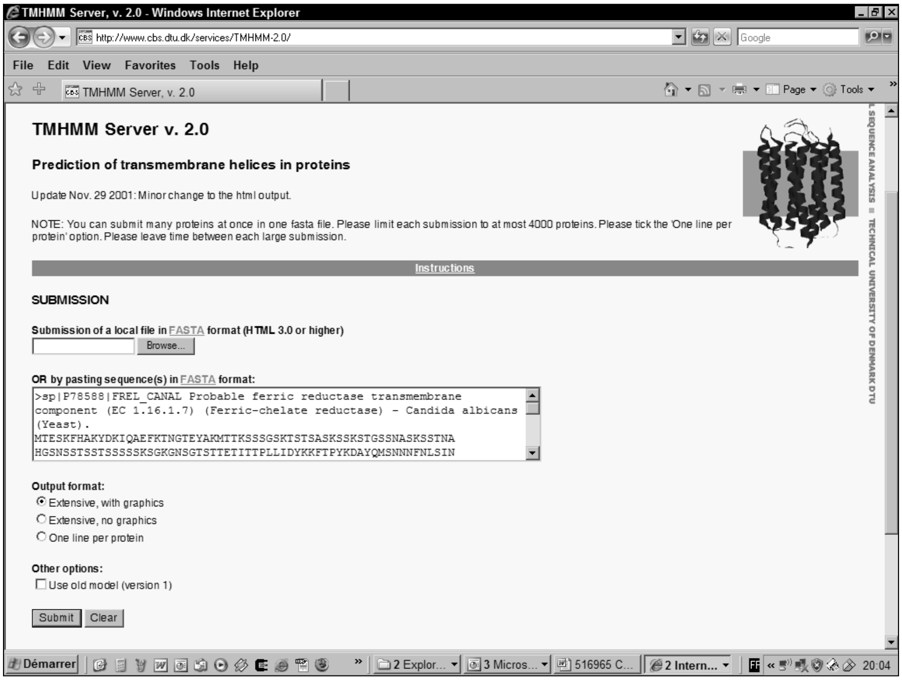
   TMHMM only recognizes the sequence in FASTA format. For the following example, we use the protein sequence FREL_CANAL. To obtain this sequence from Swiss-Prot:

   a. Open a new browser window.

   b. Open the page `www.expasy.ch/cgi-bin/get-sprot-fasta?FREL_CANAL`, which contains REL_CANAL in FASTA format.

   c. Copy the sequence onto the Clipboard.

   d. Paste the sequence in the TMHMM window (refer to Figure 6-7).



TMHMM Server, v. 2.0 - Windows Internet Explorer

http://www.cbs.dtu.dk/services/TMHMM-2.0/

TMHMM Server v. 2.0

Prediction of transmembrane helices in proteins

Update Nov. 29 2001: Minor change to the html output.

NOTE: You can submit many proteins at once in one fasta file. Please limit each submission to at most 4000 proteins. Please tick the 'One line per protein' option. Please leave time between each large submission.

Instructions

SUBMISSION

Submission of a local file in FASTA format (HTML 3.0 or higher)

OR by pasting sequence(s) in FASTA format:

```
>sp|P78588|FREL_CANAL Probable ferric reductase transmembrane
component (EC 1.16.1.7) (Ferric-chelate reductase) - Candida albicans
(Yeast).
MTESKFHAKYDKIQAEFKTNGTEYAKMTTKSSSGSKTSTSASKSSKSTGSSNASKSSTNA
HGSNSSTSSTSSSSSKSGKGNSGTSTTETITTPLLIDYKKFTPYKDAYQMSNNNFNLSIN
```

Output format:

● Extensive, with graphics
○ Extensive, no graphics
○ One line per protein

Other options:

□ Use old model (version 1)

Figure 6-7: The bottom half of the TMHMM page.

3. **Keep the Output Format radio buttons to their default value (Extensive with Graphics).**

4. **Press the Submit button.**

5. **When the Results page appears, choose File➪Save As from your browser main menu to save your results.**

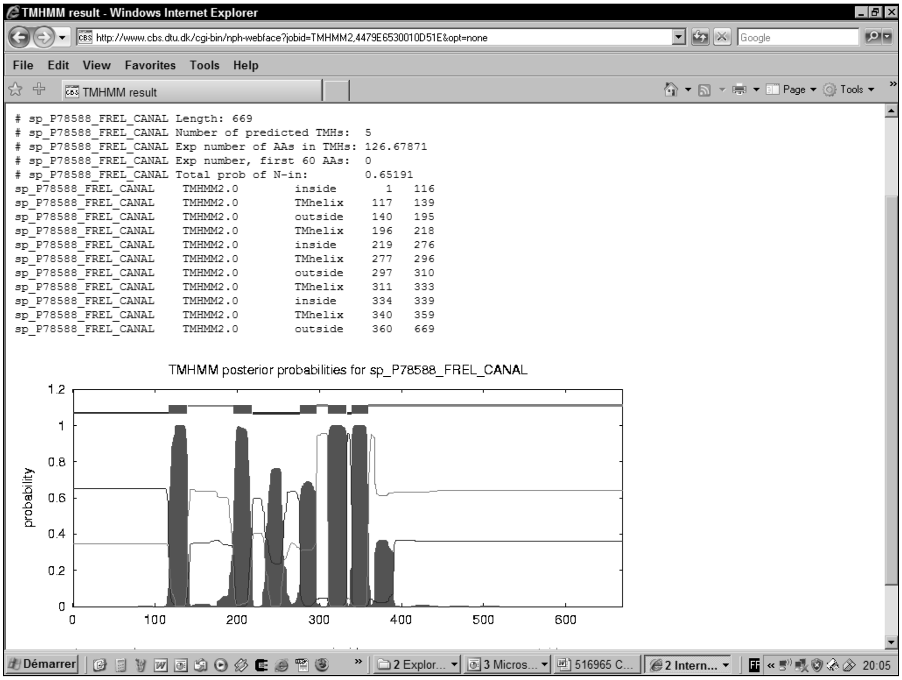
   If you are using Internet Explorer, the Save As Type pull-down menu appears; choose Web Page, Complete (`*.htm`, `*.html`). Your file is saved along with a directory that has the same name. Don't separate your file from this directory.

Interpreting results from TMHMM

Although results from TMHMM are slightly different from those you get from ProtScale, they are nevertheless in rough agreement. This is good news because the two methods use very different principles.

There are two major differences between ProtScale and TMHMM:

- TMHMM returns a precise prediction. This prediction is at the top of the output — just before the plotted graph — as shown in Figure 6-8.
- TMHMM predicts the segments that are inside the cell and the segments that are outside the cell.



TMHMM result - Windows Internet Explorer
http://www.cbs.dtu.dk/cgi-bin/nph-webface?jobid=TMHMM2,4479E6530010D51E&opt=none
Google
File Edit View Favorites Tools Help
TMHMM result
Page
Tools

```
# sp_P78588_FREL_CANAL Length: 669
# sp_P78588_FREL_CANAL Number of predicted TMHs:  5
# sp_P78588_FREL_CANAL Exp number of AAs in TMHs: 126.67871
# sp_P78588_FREL_CANAL Exp number, first 60 AAs:  0
# sp_P78588_FREL_CANAL Total prob of N-in:        0.65191
sp_P78588_FREL_CANAL    TMHMM2.0        inside       1   116
sp_P78588_FREL_CANAL    TMHMM2.0        TMhelix    117   139
sp_P78588_FREL_CANAL    TMHMM2.0        outside    140   195
sp_P78588_FREL_CANAL    TMHMM2.0        TMhelix    196   218
sp_P78588_FREL_CANAL    TMHMM2.0        inside     219   276
sp_P78588_FREL_CANAL    TMHMM2.0        TMhelix    277   296
sp_P78588_FREL_CANAL    TMHMM2.0        outside    297   310
sp_P78588_FREL_CANAL    TMHMM2.0        TMhelix    311   333
sp_P78588_FREL_CANAL    TMHMM2.0        inside     334   339
sp_P78588_FREL_CANAL    TMHMM2.0        TMhelix    340   359
sp_P78588_FREL_CANAL    TMHMM2.0        outside    360   669
```

Figure 6-8: TMHMM results on the Swiss-Prot Protein FREL_CANAL.

In this case, TMHMM fails to predict the segment that is in the middle (the one between amino acids 234 and 255), but it gives a fairly good estimation for five of the segments.

For each residue of your sequence, the graph in Figure 6-8 displays the probability of it being either in a transmembrane domain, inside the cell, or outside the cell.

This probability is only a prediction; it's not necessarily correct.

If you really need an accurate prediction because you're about to design an experiment, we recommend running many predictions using different methods. Very different methods that give you the same results are usually a good indication that you're on the right track.

Looking for coiled-coil regions

Coiled-coil regions are portions of a protein formed by the intertwining of two or three alpha-helices. One reason it's considered interesting to find coiled-coil regions is that they're often involved in protein-protein interactions. Another (less glorious) reason is that these coiled-coil regions can give false matches when you do a database search (see Chapter 7) — and it can be a good thing to filter them out. If you want to predict these regions in your protein of interest, you can use the conveniently named COILS server at EMBnet. Check them out at

```
www.ch.embnet.org/software/COILS_form.html
```

Predicting Post-Translational Modifications in Your Protein

Proteins often need to be modified before they become active in the cell. Biologists call such operations *post-translational modifications* because they occur after the translation (synthesis) of the protein. These modifications may involve adding sugars, modifying amino acids, or removing pieces of the newly synthesized protein. If you're studying a new protein, you probably want to know about such matters. This may be very important if you want to clone and express a human protein in bacteria — because, in order to be active, your protein may require some post-translational modifications that the bacterium itself cannot make.

One of the most useful tools for analyzing post-translational modifications is PROSITE — a database you can find on the ExPASy site. It contains a list of short sequence motifs (also some named patterns) that experiments have associated with particular biological properties. Many of these patterns are associated with post-translational modifications. On the ExPASy server (`www.expasy.org`), you can compare your protein sequence with the collection of patterns in PROSITE — and find out which modifications your protein is likely to undergo.

WARNING!

When you do sequence analysis, there is something you must ALWAYS remember: Similar short sequences (such as those with less than 20 amino acids) don't ALWAYS have the same function. Thus, if a small sequence has been shown to function as an ATP binding in a protein — and if the protein you're analyzing contains a short segment with EXACTLY the same sequence — it doesn't NECESSARILY mean that your protein is also an ATP binding protein. What you have is an indication that it *may* be an ATP binding protein. Of course, the longer the segment, the stronger the indication.

This warning on the meaning of short similarity regions also applies to PROSITE patterns — patterns listed there are often quite short. That said, we can show you how to check your sequence to see whether it contains any known PROSITE pattern.

Looking for PROSITE patterns

ScanProsite is a server that allows you to compare your protein with the list of patterns contained in the PROSITE database. Highly trained specialists designed each pattern in this database. If you find that your protein contains a PROSITE pattern, this one fact can (often) give you a pretty clear indication of its function.

The PROSITE patterns

When biologists first started having access to protein sequences, one of their first discoveries was that some small conserved sequences are often associated with important properties — such as cellular localization, ligand binding, and so on.

Amos Bairoch, the creator of Swiss-Prot, started exploiting this discovery while building his protein database. To help organize and annotate the proteins, he created a collection of small, well-conserved segments that he could use to classify and analyze new proteins. These special segments are known as *patterns* — and they are still widely used today as a means of characterizing new proteins. PROSITE is the name Amos gave to this particular pattern database. These days, PROSITE no longer contains only patterns; it also includes a new, more sophisticated type of model: the *profile.* Profiles describe every position of an entire protein family — not just a few highly conserved positions, as patterns do. (These domain profiles are very different from the property profiles we also describe in this chapter.)

Here's how you can get ScanProsite working for you:

1. **Point your browser to** `www.expasy.org/tools/scanprosite/`**.**

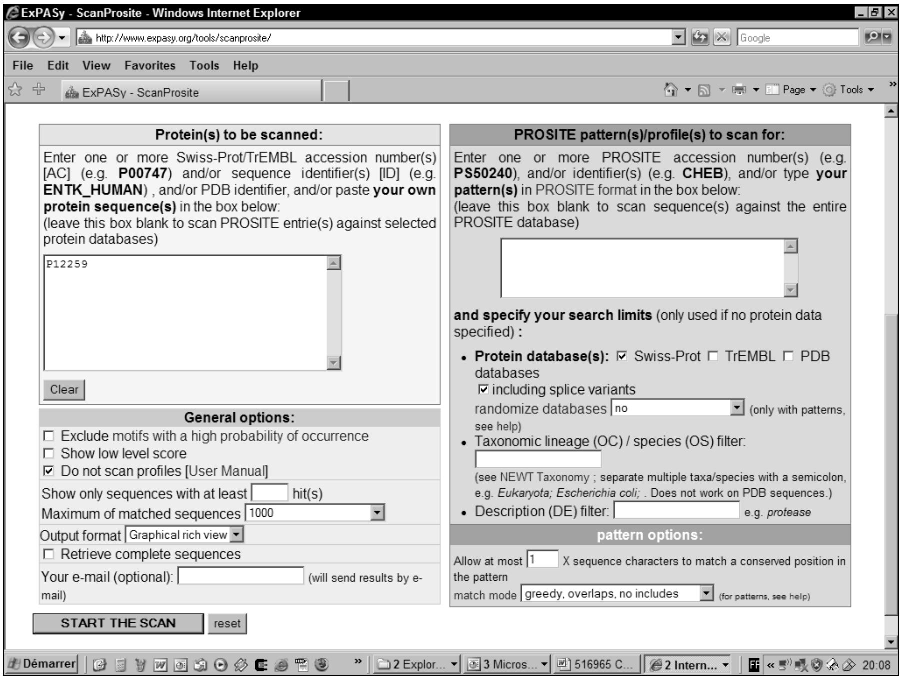
   The ScanProsite page of the ExPASy site appears. If you scroll down the page, you'll see two distinct sections:

   - The right section in Figure 6-9 is the place to go if you already have a pattern and you want to check whether other proteins in Swiss-Prot contain this pattern.
   - The left section in Figure 6-9 is right up your alley if you have a protein sequence and you want to find out which PROSITE pattern it contains.

2. **Paste your sequence or enter its accession number into the search text box on the left (in the Protein[s] to be Scanned section).**

   We want to go from a protein sequence to a pattern, so we're staying with the search box on the left. You can use

   - Raw format (amino acids only, space and numbers are tolerated but removed automatically).
   - An accession number. For this example we use the Swiss-Prot protein `P12259`, which is the precursor of the Human Coagulation factor V.



Protein(s) to be scanned:

Enter one or more Swiss-Prot/TrEMBL accession number(s) [AC] (e.g. **P00747**) and/or sequence identifier(s) [ID] (e.g. **ENTK_HUMAN**) , and/or PDB identifier, and/or paste **your own protein sequence(s)** in the box below:
(leave this box blank to scan PROSITE entrie(s) against selected protein databases)

```
P12259
```

Clear

General options:

☐ Exclude motifs with a high probability of occurrence
☐ Show low level score
☑ Do not scan profiles [User Manual]

Show only sequences with at least ____ hit(s)
Maximum of matched sequences 1000
Output format Graphical rich view
☐ Retrieve complete sequences
Your e-mail (optional): ____ (will send results by e-mail)

START THE SCAN reset

PROSITE pattern(s)/profile(s) to scan for:

Enter one or more PROSITE accession number(s) (e.g. **PS50240**), and/or identifier(s) (e.g. **CHEB**), and/or type **your pattern(s)** in PROSITE format in the box below:
(leave this box blank to scan sequence(s) against the entire PROSITE database)

and specify your search limits (only used if no protein data specified) **:**

- **Protein database(s):** ☑ Swiss-Prot ☐ TrEMBL ☐ PDB databases
  ☑ including splice variants
  randomize databases no (only with patterns, see help)
- Taxonomic lineage (OC) / species (OS) filter:
  (see NEWT Taxonomy ; separate multiple taxa/species with a semicolon, e.g. *Eukaryota; Escherichia coli;* . Does not work on PDB sequences.)
- Description (DE) filter: ____ e.g. *protease*

pattern options:

Allow at most 1 X sequence characters to match a conserved position in the pattern
match mode greedy, overlaps, no includes (for patterns, see help)

Figure 6-9: The ScanProsite Interface on the ExPASy Server.

3. **Uncheck the Exclude Motifs with a High Probability of Occurrence box.**

   By default, the server ignores small patterns that may not be statistically meaningful. If you want these small patterns to be considered, be sure you uncheck this box.

4. **Check the Do Not Scan Profiles box.**

   For the purposes of this steps list, it's okay not to select to scan the profiles; that takes more time. Furthermore, when it comes to post-translational modifications, patterns are usually more interesting than profiles.

5. **Press the Start the Scan button and wait.**

   An intermediate page appears. Keep waiting until your browser displays a Results page like the one shown in Figure 6-10.

   Wait for the results! Don't click any of the hyperlinks that appear on the intermediate page. The ScanProsite server can be very slow at some times of the day.

   If your results never arrive, you have the alternative of using any ExPASy site that is closer to you or running in a part of the world that is sleeping. (See the listing at the end of this chapter for alternative ExPASy addresses.)

6. **Choose File⇨Save As from your browser's main menu to save the information on the Results page.**

Interpreting ScanProsite results

The main problem with most post-translational patterns is that they are short. The short sequences these patterns match may occur by chance, and could have no real biological significance.

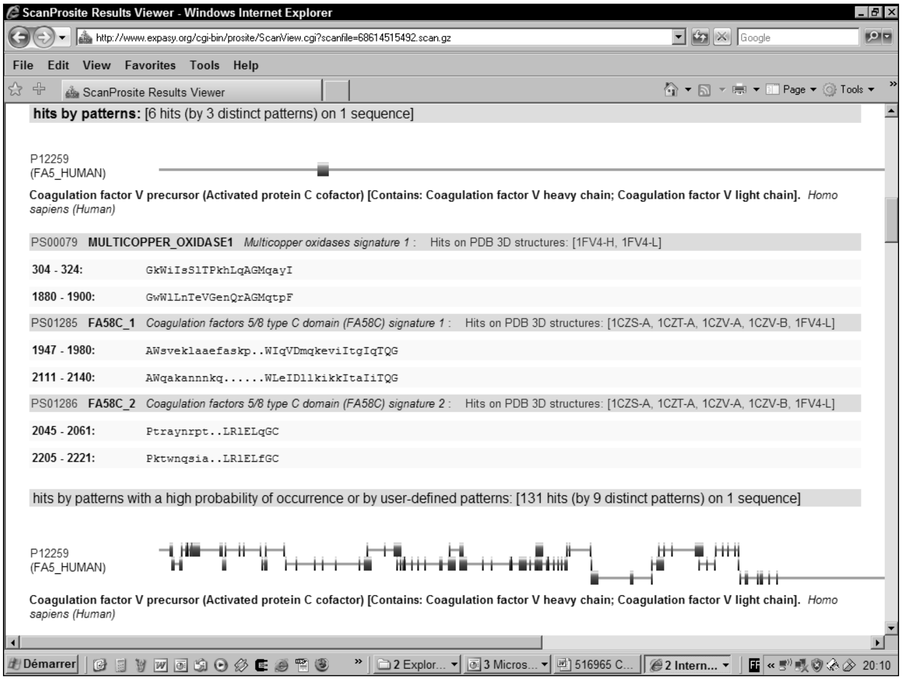
The rule is that similarities between short sequences must always be treated with suspicion. A good way to rule out many unlikely modifications is to read carefully the documentation associated with the pattern (the PDOC).

Understanding the ScanProsite output

The first section of the output is named "Hits by Patterns" and shows the location of every type of pattern found on your protein. It is basically a summary where each pattern family has its own color code. Pass the mouse pointer over one of the colored rectangles to see its name displayed.

In moving down through the Hits by Patterns section, you can see the details of each pattern found in your sequence. Documentation is available via the hyperlink displayed in the output. A portion of this output is shown in Figure 6-10. For each pattern found in PROSITE, ScanProsite reports the following information:

- **PS#####:** The first hyperlink leads you to the pattern documentation, a high-quality documentation that contains a lot of relevant information on your pattern and its potential biological function.
- **The pattern name, in bold characters.**
- **Structure:** Some patterns have been identified in proteins with a known 3-D structure. If this is the case with the pattern you're looking at, you can find the PDB name of the structure(s) here — as well as a hyperlink to the 3-D representation of this structure. PDB stands for the *P*rotein *D*ata *B*ase, a database that contains all the known protein 3-D structures. (See Chapter 11 for more information on the PDB.) If you click one of these hyperlinks, a static GIF image of the corresponding structure appears in your browser. Figure 6-11 shows the image you can obtain by clicking the 1czv hyperlink (or use the address `www.expasy.org/cgi-bin/view-pdb?pdb=1czv&ps=PS01285`). The image contains
  - A 3-D structure of your protein using a ribbon representation
  - The location of this particular pattern (marked in green) in the structure as a whole
- **The list:** A list of the segments that contain the patterns within your protein. The numbers indicate the position of the match within your sequence. Capital letters indicate residues that were specified by the pattern; lowercase letters indicate residues that weren't specified by the pattern.


ScanProsite Results Viewer - Windows Internet Explorer
http://www.expasy.org/cgi-bin/prosite/ScanView.cgi?scanfile=68614515492.scan.gz
Google
File Edit View Favorites Tools Help
ScanProsite Results Viewer
Page
Tools
Démarrer
2 Explor...
3 Micros...
516965 C...
2 Intern...
20:10

hits by patterns: [6 hits (by 3 distinct patterns) on 1 sequence]

P12259
(FA5_HUMAN)

Coagulation factor V precursor (Activated protein C cofactor) [Contains: Coagulation factor V heavy chain; Coagulation factor V light chain]. *Homo sapiens (Human)*

PS00079 **MULTICOPPER_OXIDASE1** *Multicopper oxidases signature 1 :* Hits on PDB 3D structures: [1FV4-H, 1FV4-L]

304 - 324: `GkWiIsSlTPkhLqAGMqayI`

1880 - 1900: `GwWlLnTeVGenQrAGMqtpF`

PS01285 **FA58C_1** *Coagulation factors 5/8 type C domain (FA58C) signature 1 :* Hits on PDB 3D structures: [1CZS-A, 1CZT-A, 1CZV-A, 1CZV-B, 1FV4-L]

1947 - 1980: `AWsveklaaefaskp..WIqVDmqkeviItgIqTQG`

2111 - 2140: `AWqakannnkq......WLeIDllkikkItaIiTQG`

PS01286 **FA58C_2** *Coagulation factors 5/8 type C domain (FA58C) signature 2 :* Hits on PDB 3D structures: [1CZS-A, 1CZT-A, 1CZV-A, 1CZV-B, 1FV4-L]

2045 - 2061: `Ptraynrpt..LRlELqGC`

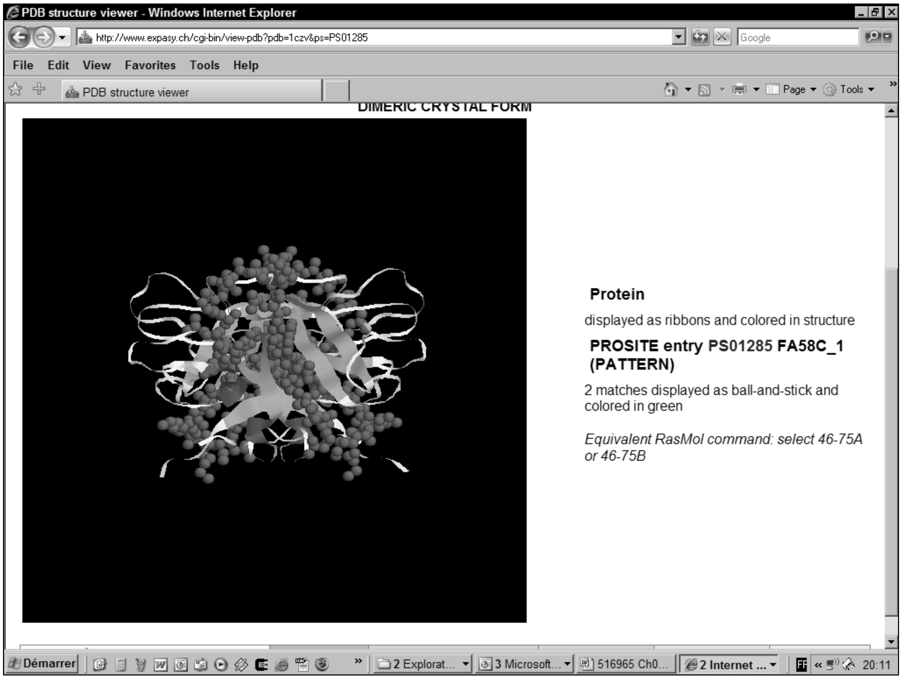
2205 - 2221: `Pktwnqsia..LRlELfGC`

hits by patterns with a high probability of occurrence or by user-defined patterns: [131 hits (by 9 distinct patterns) on 1 sequence]

P12259
(FA5_HUMAN)

Coagulation factor V precursor (Activated protein C cofactor) [Contains: Coagulation factor V heavy chain; Coagulation factor V light chain]. *Homo sapiens (Human)*

Figure 6-10: The ScanProsite Output.


PDB structure viewer - Windows Internet Explorer
http://www.expasy.ch/cgi-bin/view-pdb?pdb=1czv&ps=PS01285
Google
File Edit View Favorites Tools Help
PDB structure viewer
Page Tools
Démarrer
2 Explorat...
3 Microsoft...
516965 Ch0...
2 Internet ...
20:11

DIMERIC CRYSTAL FORM

Protein

displayed as ribbons and colored in structure

PROSITE entry PS01285 FA58C_1 (PATTERN)

2 matches displayed as ball-and-stick and colored in green

Equivalent RasMol command: select 46-75A or 46-75B

Figure 6-11: Localization of a PROSITE pattern on the 3-D structure of a protein.

The next section of the output (titled "Hits by Patterns with a High Probability of Occurrence") is similar to this one, and shows you the location of short common patterns in your sequence.

Being careful with short patterns

The complete output obtained from P12259 gives us a good illustration of the irrelevance of most short patterns: This output indicates that 19 sites of myristillation exist in the Human Coagulation Factor V. Is this information you can trust or not?

If you click on the link PS0008 near the end of the "Hits by Patterns with a High Probability of Occurrence" section, it will take you to the documentation where you will discover that myristate is a fatty acid attached to the N-terminus (left end of the sequence) of a protein, anchoring it into the membrane. None of the hits reported here are on the N-terminus. So you can safely conclude that none of these matches are genuine.

Using the species information

When you find a post-translational modification, make sure that this modification is consistent with the species the sequence comes from. The modifications that occur in prokaryotes are usually different from those that occur in eukaryotes. You must look for this information in the PDOC corresponding to your pattern. For instance, myristillation only happens in eukaryotes. If you find a myristillation site in a protein that comes from a prokaryote, you can safely ignore it.

Weak signals add up to give strong signals

Further along the output that ScanProsite returns, the pattern list contains two matches that are specific of coagulation factors: `FA58C_1` and `FA58C_2`.

If you look at them individually, neither pattern is really significant. On the other hand, the fact that you find two related patterns at so close a distance is very exciting: It's a good indication that both patterns are probably genuine.

This is a good example of the fact that two weak indications can make up for strong evidence when they are consistent (and independent!). It's a bit like two shortsighted witnesses telling you exactly the same thing about what they saw. For an investigator, this is even more interesting than the report of one eagle-eye witness, as long as the investigator can be sure that the two shortsighted witnesses never had a chance to talk with each other.

Eliminating weak patterns

When you have too many weak patterns, a good strategy is to build a multiple sequence alignment with related sequences. A pattern that corresponds to a genuine post-translational modification is normally well conserved. We provide you with a brief session on multiple alignments in Chapter 2 and most of what you need to know to build high-quality multiple sequence alignments is in Chapter 9.

Everything is not in PROSITE

Also important to know is that the PROSITE pattern collection doesn't contain a description for every known post-translational modification. If you can't find what you need in PROSITE, the ExPASy server gives you links to other tools that are specialized for this type of analysis. (The links are located in the Post-Translational Modification Prediction section of PROSITE at `www.expasy.org/tools/#ptm`.)

Finding Known Domains in Your Protein

Gurus define domains as "independent globular folding units." For the rest of us, a *domain* is a portion of protein that can keep its shape if you remove it from the rest of the protein. A domain consists of at least 50 amino acids. Domains are like the various components of your kitchen: the oven, the microwave, the fridge, and so on. All together they constitute the complete kitchen, but they can also exist separately: You're allowed to remove the microwave from your kitchen if you just have to have that late-night popcorn in your lab.

Your average protein consists of two or three domains. Figure 6-12 is a snapshot of the TolB, a protein that is made up of two domains. Usually, each domain plays a specific role in the function of the protein. It may interact with other proteins, it may bind an ion like calcium or zinc, or it may contain an active site. It is common to have a catalytic domain associated with a binding domain and a regulatory domain. Imagine, if you will, a toaster, where you have the grill (catalytic), the toast holder (binding), and the switch (regulation).

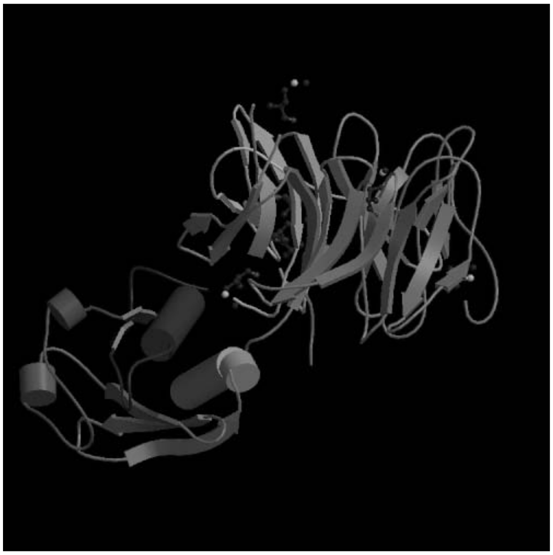


Figure 6-12: The TolB protein is made of two domains.

In principle and in practice, little difference lies between searching patterns and searching domains in your sequence, and once a domain is characterized, it is easy to check whether your protein contains it or not. A dozen or so domain collections exist around the world (Table 6-1). For some of these collections, a bunch of specialists handcrafted the domains one by one, like Swiss watches. These collections give very precise results but are sometimes a bit incomplete. (See the entries marked Manual in Table 6-1.) In other collections, a computer crunches all known sequences and splits them into a bunch of domains. These collections are more extensive but less informative. (See the entries marked Automatic in Table 6-1.)

Choosing the right collection of domains

Domains are so useful and important in biology that biologists have been trying to build nice collections of them for quite some time already. The problem is twofold: (a) it can be tricky to define a specific domain unambiguously, and (b) gurus often find it hard to agree with each other. (So now you know why various authors have described most of the important domains in slightly different ways.)

For us, the consequences are that there's a large redundancy among the eight major collections of domains available today (Table 6-1). Life would be simpler if we could tell you that TastyDom (r) is the best and that you must not use PooPooDom (r)! Unfortunately, this isn't the way it works. Each collection has its pros and cons — and eventually, if you really want to understand your protein, you need to look at each collection! It takes time, but in the end, it can be truly worth the effort.

The InterPro project has made it possible to search many domain databases simultaneously. In Table 6-1, (IP) indicates the domain databases that can be accessed by InterPro. InterPro helps you make sure that you're not missing even the tiniest bit of information. You can also find out easily when some domain collections disagree on your sequence.

Unfortunately, InterPro is not exhaustive. The only way to make complete analyses of the domains contained in your sequence is to use the three major domain servers:

- InterProScan
- CD-Search
- Motif-Scan

It is highly probable that one, and *only* one, of these servers contains *the* bit of information that helps you make sense of your protein — meaning that you'll have to check out each server in turn. The rest of this chapter shows you how to use each of these servers and get the best out of each.

Table 6-1 The Main Domain Collections

Name	*Web Address*	*Number of Domains*	*Generation*
PROSITE-Profile (IP)	`www.expasy.org/prosite`	616	Manual
PfamA (IP)	`www.sanger.ac.uk/Software/Pfam`	7973	Manual

Name	*Web Address*	*Number of Domains*	*Generation*
PRINTs (IP)	`www.bioinf.man.ac.uk/dbbrosers/PRINTS`	1900	Manual
PRODOM (IP)	`protein.toulouse.inra.fr/prodom/current/html/home.php`	736000	Automatic
SMART (IP)	`smart.embl-heidelberg.de`	685	Manual
COGs	`www.ncbi.nlm.nih.gov/COG/new/`	4852	Manual
TIGRFAM (IP)	`www.tigr.org/TIGRFAMs`	2453	Manual
BLOCKs	`blocks.fhcrc.org/`	12542	Automatic

Finding domains with InterProScan

You want to use the InterProScan server if you have a protein sequence and you want to know which domain this sequence may contain. InterProScan allows you to compare your sequence with InterPro, a domain database that includes most of the major domain collections available online. In a way, InterProScan is for protein sequences and domains what metasearch engines are for the World Wide Web!

Here's how you can get InterProScan working for you:

1. **Point your browser to** `www.ebi.ac.uk/InterProScan/`**.**

   The InterProScan home page appears.

2. **Enter your sequence in the search text box.**

   Most legal formats are recognized here (including the raw format), but the accession numbers don't work.

   In the following example, we use the protein sequence FOSB_HUMAN. Here's how to obtain this sequence from Swiss-Prot:

   a. Open a new Internet browser window.

   b. Open the page `www.expasy.ch/cgi-bin/get-sprot-fasta?FOSB_HUMAN`, which contains FOSB_HUMAN in format FASTA.

c. Copy the sequence.

d. Paste your sequence in the InterProScan search text box, as shown in Figure 6-13.

You can also scan a nucleotide sequence. If you want to do so, select DNA instead of PROTEIN in the selector just above the sequence.

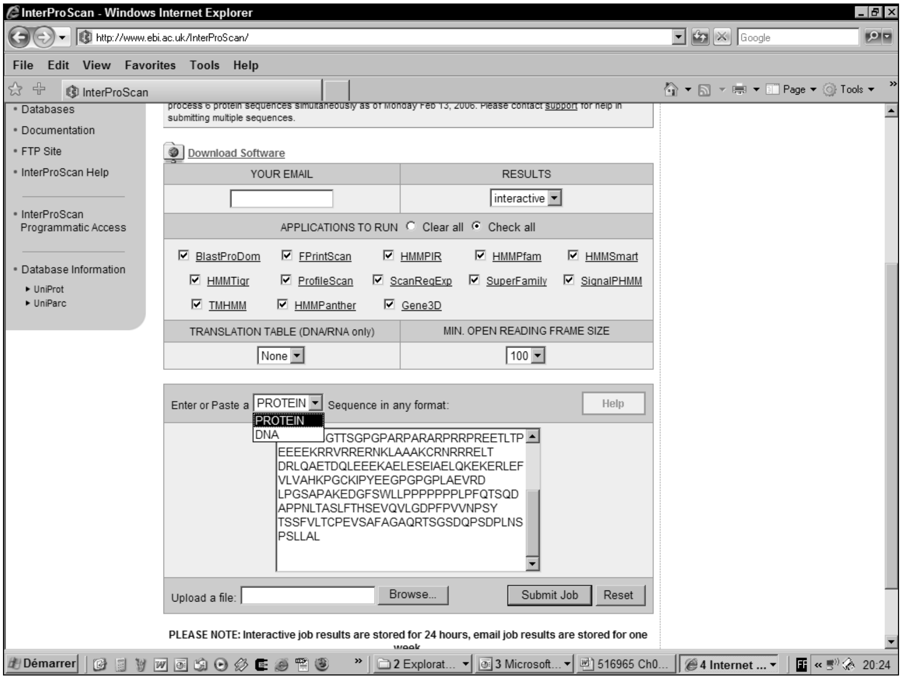
3. **In the Applications to Run section, choose the domain databases you are interested in.**

Interpro lets you search domain databases, but it also lets you run some specific prediction applications, such as *TMHMM* (for trans-membrane predictions) and *SignalPHMM,* which runs a prediction to determine whether your sequence contains a *signal peptide* (a short N-terminus segment that causes your protein to reach its destination in the cell, as precisely as a Zip code.)

You can obtain your results faster if you remove some of the largest databases (ProDom, for example) from your search.

4. **Click the Submit Job button.**

An intermediate page pops up and informs you of how long your search has been running.



process 6 protein sequences simultaneously as of Monday Feb 13, 2006. Please contact support for help in submitting multiple sequences.

YOUR EMAIL	RESULTS
	interactive

APPLICATIONS TO RUN ○ Clear all ● Check all

☑ BlastProDom ☑ FPrintScan ☑ HMMPIR ☑ HMMPfam ☑ HMMSmart
☑ HMMTigr ☑ ProfileScan ☑ ScanRegExp ☑ SuperFamily ☑ SignalPHMM
☑ TMHMM ☑ HMMPanther ☑ Gene3D

TRANSLATION TABLE (DNA/RNA only)	MIN. OPEN READING FRAME SIZE
None	100

Enter or Paste a PROTEIN Sequence in any format:

PROTEIN
DNA

```
GTTSGPGPARPARARPRRPREETLTP
EEEEKRRVRRERNKLAAAKCRNRRRELT
DRLQAETDQLEEEKAELESEIAELQKEKERLEF
VLVAHKPGCKIPYEEGPGPGPLAEVRD
LPGSAPAKEDGFSWLLPPPPPPPLPFQTSQD
APPNLTASLFTHSEVQVLGDPFPVVNPSY
TSSFVLTCPEVSAFAGAQRTSGSDQPSDPLNS
PSLLAL
```

Upload a file:

PLEASE NOTE: Interactive job results are stored for 24 hours, email job results are stored for one week

Figure 6-13: The Inter ProScan home page.

Do not click any button while you wait; the results are automatically displayed when they're ready, using SRS (the Sequence Retrieval System).

5. **Save your results.**

   To save the nice color display, the easiest thing to do is to press the Prt Sc (Print Screen) key on your keyboard, and then cut and paste (Ctrl-V) the display into your favorite application (Powerpoint, for instance). To keep your exact results for further reference, the simplest thing to do is to save the raw output:

   a. Click the handy Raw Output button.

      A new window appears, displaying just the raw data.

   b. Use your browser's File➪Save As command.

Interpreting InterProScan results

Because InterProScan returns a large amount of information, interpreting its results correctly is something of an art form. Before clicking any hyperlink, make sure that you understand exactly what the result page contains.

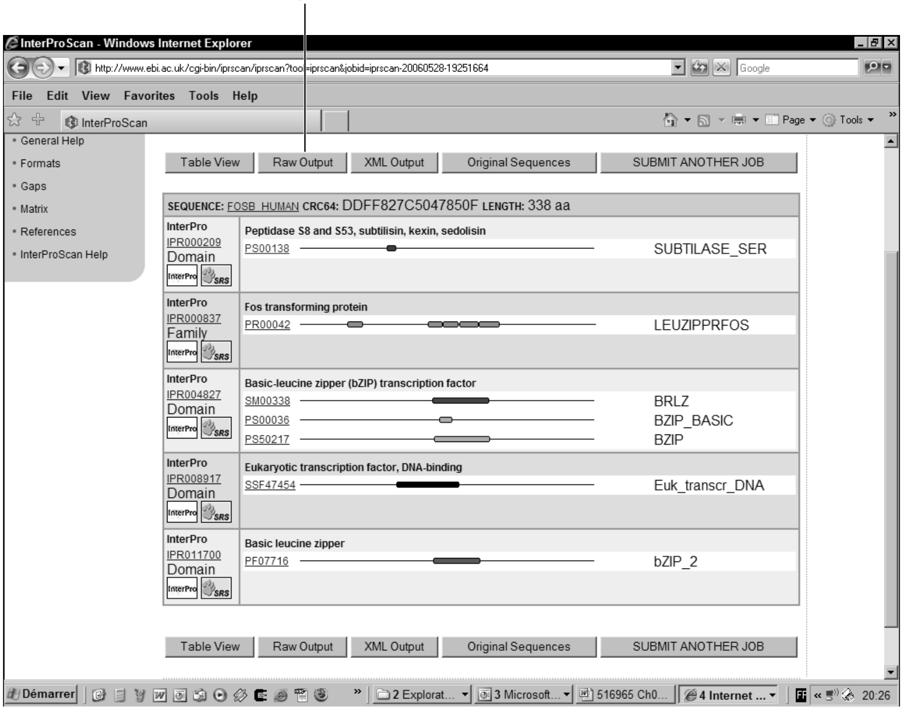
Understanding the InterProScan output

Figure 6-14 shows the results you obtain when scanning the protein FOSB_ HUMAN with the InterProScan server. It illustrates well the variety of results one can expect when you compare a sequence with domain databases.

In the output, each line represents a match with some type of InterPro domain.

- **The first entry in each column indicates the type of diagnosis provided: Family or Domain.** Some domains or signatures are specific of a complete protein family while others are simply specific of a domain. When several domains from different domain databases describe the same thing, the InterPro database groups them in the same box, like the IPR 004827 domain.
- **The IPR#### hyperlink points to the InterPro documentation.** Here InterPro attempts to summarize the information spread in the documentations of various domain databases. It's always a good idea to read the InterPro documentation, but don't stop there — be sure to read the documentation for the individual sequences as well.
- **In front of each line is a hyperlink that can take you to the domain entry in its database.** For instance, if you click the PS00138 link, your browser takes you to the entry 138 of the PROSITE database, where you can find the individual PROSITE documentation.
- **On every line, little colored boxes show you where the match occurred on your sequence.** Their size is proportional to the size of the domain. If you pass the mouse pointer over these rectangles, the coordinates of the match (in your sequence) appear on-screen.

Raw Output button



SEQUENCE: FOSB_HUMAN CRC64: DDFF827C5047850F LENGTH: 338 aa

InterPro	Description	Match	Name
InterPro IPR000209 Domain	Peptidase S8 and S53, subtilisin, kexin, sedolisin	PS00138	SUBTILASE_SER
InterPro IPR000837 Family	Fos transforming protein	PR00042	LEUZIPPRFOS
InterPro IPR004827 Domain	Basic-leucine zipper (bZIP) transcription factor	SM00338 PS00036 PS50217	BRLZ BZIP_BASIC BZIP
InterPro IPR008917 Domain	Eukaryotic transcription factor, DNA-binding	SSF47454	Euk_transcr_DNA
InterPro IPR011700 Domain	Basic leucine zipper	PF07716	bZIP_2

Figure 6-14: Output of the Inter ProScan Server.

Making sense of the inconsistencies in the InterProScan output

You can see in Figure 6-14 that all domain databases agree on the presence of a series of Leucine zippers that are roughly in the middle of the sequence.

On the other hand, the domain databases do *not* agree *exactly* on the boundaries of these matches. Deciding who is right and who is wrong is always a bit of an art, but in this case a majority vote could be a good indication. (Now you know why using all the domain collections — and not just one — is so important.)

TIP

Another reason why seven databases are better than one is that there's always a good chance that one of the databases may contain a domain that isn't in the others. This is especially true with exotic domains that are not yet included in all domain collections.

Of course, the fact that a domain is or is not reported isn't 100 percent proof of anything. (See the sidebar "A common mistake when scanning domains.") To make sure that a reported hit is genuine, we recommend taking a close look at the alignment between the profile and your sequence. Unfortunately, InterProScan doesn't let you do this, which is why you also need to use the other servers (such as CD-Search and Motif-Scan) that we describe in the next section.

Finding domains with the CD server

The CD (Conserved Domain) server of the NCBI follows the same principle as the InterProScan. The main advantage of the CD server is that reported hits come with a score that helps you discriminate the good from the spurious matches. On the downside, the CD server doesn't integrate as many databases as InterProScan, although it contains quite a few domains contributed by the NCBI that you can't find anywhere else.

If you're interested in giving the CD server a try, here's how you'd go about it:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi`**.**

   You can also access this server from the main NCBI BLAST page. Point your browser to `www.ncbi.nlm.nih.gov/BLAST/` and click the Search the Conserved Domain Database using RPS-BLAST link. This hyperlink is the fourth one in the second column.

A common mistake when scanning domains

Don't forget that when a domain server reports some hits between your sequence and a domain, this information always results from an alignment between the domain (profile) you've found and the sequence you're using for comparison. If the program believes the alignment is good enough, it will report the match. Otherwise it will not.

Unfortunately, neither the profile nor the programs are perfect — and mistakes occur. The server can tell you that your protein contains a specific domain when that isn't true, or it may fail to report a domain that *is* present in your protein. If you trust the results blindly, there's always a chance you may get it wrong.

The InterProScan server isn't very helpful when it comes to avoiding that type of mistake: It doesn't report the score of the hits and does not even display the alignments. The other servers we show you in this chapter (CD-Search and Motif Scan) give you a score along with the hits. This score has a statistical meaning, and it informs you how likely it is that your match may have occurred by chance only and is devoid of a biological meaning.

Conservative interpretations (that is, only believing very high scores) are almost always correct. On the other hand, if the score isn't so good — or if only a portion of the domain matches your sequence — you need additional evidence to make sure that what you see is real.

2. **Paste your sequence or its identifier in the large box near the top.**

   For this example, we chose the protein FOSB_HUMAN (P53539). If you want to use this sequence: Type **P53539** in the Sequence box, as shown in Figure 6-15.

3. **Deselect the Apply Low Complexity Filter check box, located under the Advanced Search Options heading.**

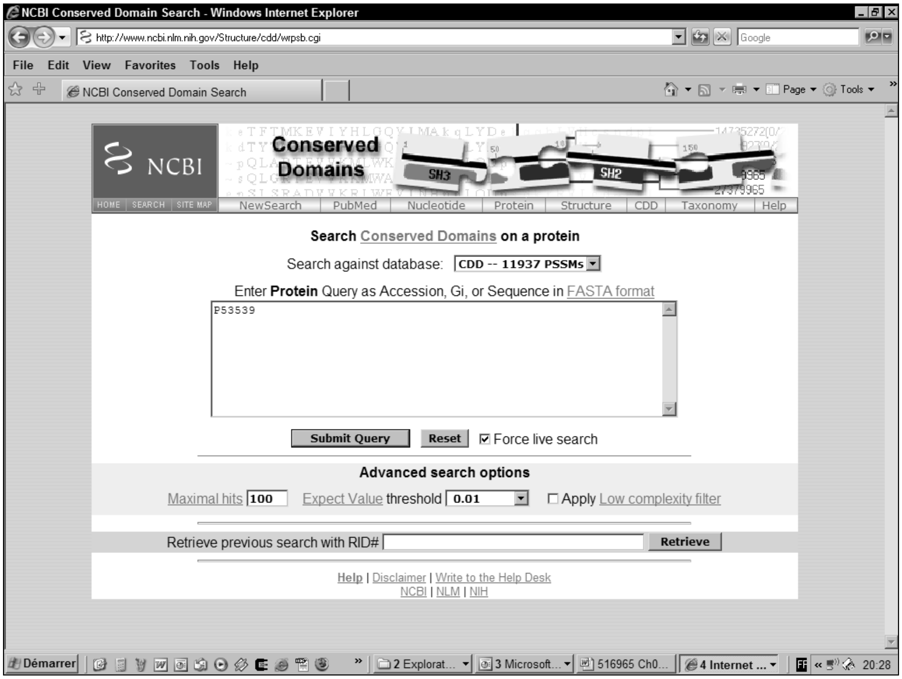
   In many interesting domains, a particular type of amino acid is overrepresented. For instance, there are more leucines than expected by chance in the leucine zippers, or more glycines than expected by chance in the glycine-rich domains — and so on — for many domains.

   Because repeated residues make a sequence simpler, sequences that contain them are described as *low-complexity*. If you keep the Apply Low Complexity Filter box checked, you may lose these domains.

4. **Set the Expect Value Threshold to 1.**

   The default value (0.01) can be too stringent, especially when considering low complexity domains.

5. **Click the Submit Query button.**


NCBI Conserved Domain Search - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Google
File Edit View Favorites Tools Help
NCBI Conserved Domain Search
Page ▾ Tools ▾
NCBI
Conserved Domains
SH3
SH2
HOME | SEARCH | SITE MAP
NewSearch | PubMed | Nucleotide | Protein | Structure | CDD | Taxonomy | Help
Démarrer
2 Explorat...
3 Microsoft...
516965 Ch0...
4 Internet ...
20:28

Search Conserved Domains on a protein

Search against database: CDD -- 11937 PSSMs

Enter **Protein** Query as Accession, Gi, or Sequence in FASTA format

```
P53539
```

Submit Query | Reset | ☑ Force live search

Advanced search options

Maximal hits 100 Expect Value threshold 0.01 □ Apply Low complexity filter

Retrieve previous search with RID# | Retrieve

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Figure 6-15: The CD search server at NCBI.

6. **Save your results.**

   The easiest way to save your graphic display is to do a screen dump by pressing Prt Sc key, and then cutting and pasting your result into a PowerPoint presentation.

Interpreting and understanding CD server results

The principle of the output from the NCBI CD server is very similar to that of InterPro, and is divided into two sections:

- **The Graphic:** The Graphic display, as shown in Figure 6-16, shows the regions of your protein that match a domain:

  - Red domains are from SMART.
  - Ragged ends indicate partial matches.

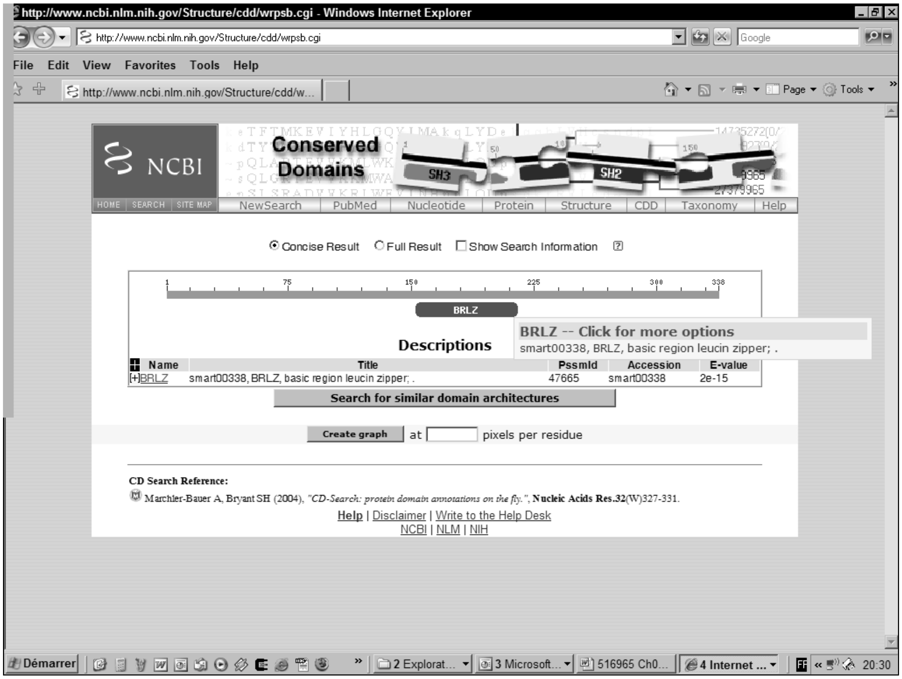
  If you pass the mouse pointer over a domain, its complete name appears in the little window along with its score. The score is an *E-value:* It tells you how many times you can expect a hit that good by sheer chance only, given your sequence and the database you scanned.

  You can interpret your score by using the following rules:

  - The lower the E-value, the better the score — and the more likely your hit is to be genuine.
  - E-values need to be below 0.01 to mean something.
  - The E-value may need to be much lower than 0.01 if you remove the low-complexity filtering (Step 3).
  - Treat ragged-end matches with suspicion, especially if they occur in low-complexity regions. In our example, none of the ragged-end matches is significant.

- **The Hit List:** The hit list (as shown in Figure 6-17) reports the domains that match your sequence, sorted by E-value. The best hits come first, shown at the top. If you click the hyperlink, a page pops up that contains the documentation associated with this entry.

  Clicking the (+) sign next to the hit will extend it into an alignment between your query and some estimated consensus of the domain sequences. This consensus is something like the average domain sequence.


http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi
Google
File Edit View Favorites Tools Help
http://www.ncbi.nlm.nih.gov/Structure/cdd/w...
Page Tools
NCBI
Conserved Domains
SH3
SH2
HOME SEARCH SITE MAP
NewSearch PubMed Nucleotide Protein Structure CDD Taxonomy Help
1 75 150 225 300 338
BRLZ
Create graph
Search for similar domain architectures
Démarrer
2 Explorat...
3 Microsoft...
516965 Ch0...
4 Internet ...
20:30

● Concise Result ○ Full Result □ Show Search Information

BRLZ -- Click for more options
smart00338, BRLZ, basic region leucin zipper; .

Descriptions

Name	Title	PssmId	Accession	E-value
[+]BRLZ	smart00338, BRLZ, basic region leucin zipper; .	47665	smart00338	2e-15

Create graph at □ pixels per residue

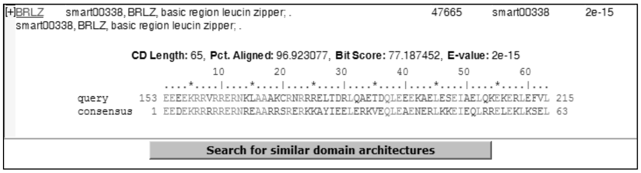
CD Search Reference:
Marchler-Bauer A, Bryant SH (2004), *"CD-Search: protein domain annotations on the fly."*, **Nucleic Acids Res.32**(W)327-331.

Help | Disclaimer | Write to the Help Desk
NCBI | NLM | NIH

Figure 6-16: Graphic output from the NCBI CD server.

If you click the Search For Similar Domain Architectures button at the bottom of the hit list, the server will fetch all the protein sequences that contain the same domains as your query sequence. These could be very remote homologues, impossible to detect in another way.

Figure 6-17: The Hit List and Alignments output from the NCBI CD server.



[+]BRLZ	smart00338, BRLZ, basic region leucin zipper; .	47665	smart00338	2e-15

smart00338, BRLZ, basic region leucin zipper; .

CD Length: 65, **Pct. Aligned:** 96.923077, **Bit Score:** 77.187452, **E-value:** 2e-15

```
                    10        20        30        40        50        60
              ....*....|....*....|....*....|....*....|....*....|....*....|...
query     153 EEEEKRRVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVL 215
consensus   1 EEDEKRRRRRERNREAARRSRERKKAYIEELERKVEQLEAENERLKKEIEQLRRELEKLKSEL 63
```

Finding domains with Motif Scan

The people who developed Motif Scan are the same folks who maintain PROSITE. The Motif Scan server provides you with the most powerful interface available for PROSITE. Motif Scan also includes some domains that have not yet been released *officially* via InterPro.

If you found nothing on the InterProScan server or on the CD server, then you've come to the right place. Motif Scan is your last chance! Here's how to take advantage of it:

1. **Point your browser to** `myhits.isb-sib.ch/cgi-bin/motif_scan`**.**

   The Motif Scan home page appears.

2. **Paste your sequence or its ID number into the Protein Sequence Input text box, as shown in Figure 6-18.**

   You have the choice among the following formats:

   - Sequence ID
   - Raw format (Residues only, space and numbers tolerated)
   - FASTA

   Motif Scan automatically recognizes the chosen format.

   For this example, you can use the FOSB_HUMAN. To do so, type **P53539** into the Protein Sequence Input box.

3. **In the Database of Motifs section, select the domain collection you want.**

   PROSITE is much smaller than Pfam. Selecting PROSITE gives you a much quicker answer than if you also select Pfam.

4. **Click the Search Button.**

   Doing so gets you a Results page (eventually).

5. **Print your results.**

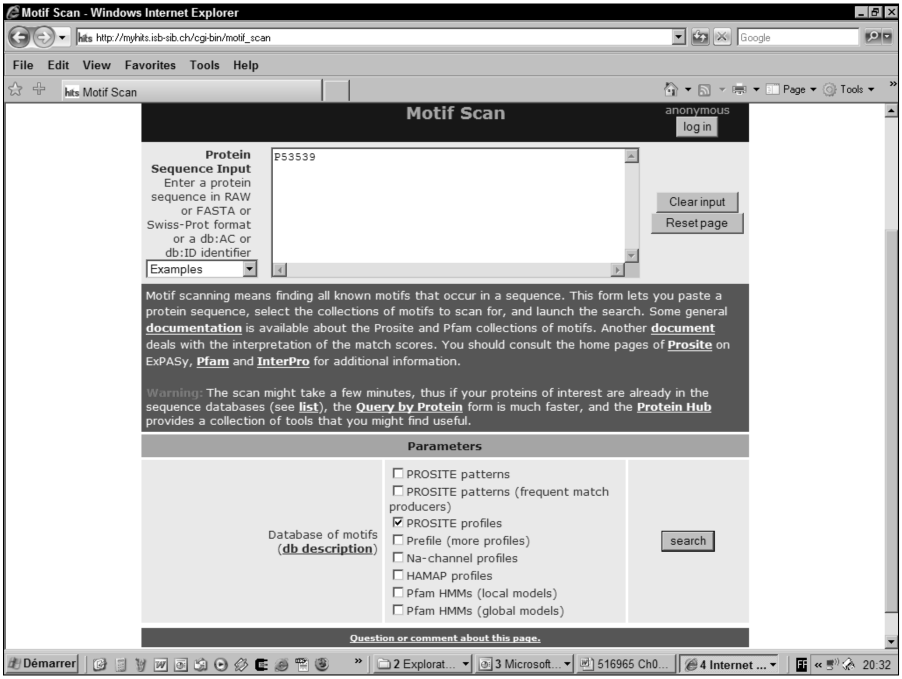
   - **Saving the graphic display:** Do a screen dump by pressing the Prt Sc key, then cut and paste it into a PowerPoint presentation.
   - **Saving your results:** Cut and paste the List of Matches section.

Interpreting and understanding the Motif Scan results

Motif Scan provides one of the richest outputs available for domain analysis. This comes at the cost of an interface that is slightly more difficult to interpret than that provided by CD or InterProScan. (See Figure 6-20)

Here are some things to keep in mind when interpreting Motif Scan output:

- **Match Map:** This match map indicates the location of every domain match on your sequence. Each match is numbered, and a legend is given at the bottom of the match.



Motif Scan

Protein Sequence Input Enter a protein sequence in RAW or FASTA or Swiss-Prot format or a db:AC or db:ID identifier

Examples

P53539

Motif scanning means finding all known motifs that occur in a sequence. This form lets you paste a protein sequence, select the collections of motifs to scan for, and launch the search. Some general **documentation** is available about the Prosite and Pfam collections of motifs. Another **document** deals with the interpretation of the match scores. You should consult the home pages of **Prosite** on ExPASy, **Pfam** and **InterPro** for additional information.

Warning: The scan might take a few minutes, thus if your proteins of interest are already in the sequence databases (see **list**), the **Query by Protein** form is much faster, and the **Protein Hub** provides a collection of tools that you might find useful.

Parameters

Database of motifs (**db description**)

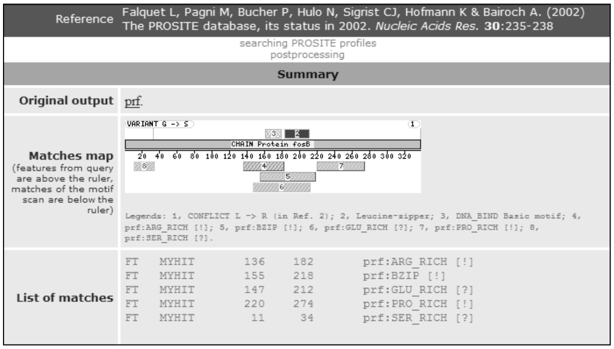
□ PROSITE patterns
□ PROSITE patterns (frequent match producers)
☑ PROSITE profiles
□ Prefile (more profiles)
□ Na-channel profiles
□ HAMAP profiles
□ Pfam HMMs (local models)
□ Pfam HMMs (global models)

Question or comment about this page.

Figure 6-18: Motif Scan server at the Swiss Institute of Bioinformatics.

- **List of Matches:** This is a text version of the match map. Copy it for further reference.
- **The Match Details Section:** Motif Scan uses the normalized score:
  - High score means a good match.
  - Only scores above 7 are considered good.

Motif Scan doesn't sort the hits according to their scores, so the best matches aren't necessarily at the top of the list. Relevant hits are indicated with an exclamation mark (`!`).



Reference Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K & Bairoch A. (2002) The PROSITE database, its status in 2002. *Nucleic Acids Res.* **30**:235-238

searching PROSITE profiles
postprocessing

Summary

Original output prf.

Matches map (features from query are above the ruler, matches of the motif scan are below the ruler)

Legends: 1, CONFLICT L -> R (in Ref. 2); 2, Leucine-zipper; 3, DNA_BIND Basic motif; 4, prf:ARG_RICH [!]; 5, prf:BZIP [!]; 6, prf:GLU_RICH [?]; 7, prf:PRO_RICH [!]; 8, prf:SER_RICH [?].

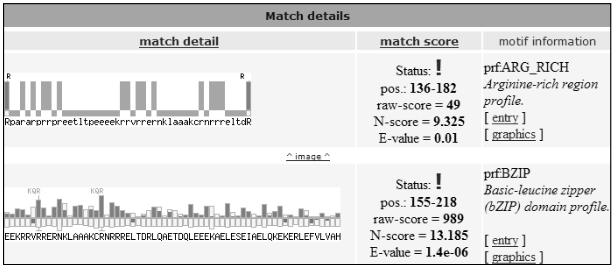
List of matches

```
FT   MYHIT       136    182        prf:ARG_RICH [!]
FT   MYHIT       155    218        prf:BZIP [!]
FT   MYHIT       147    212        prf:GLU_RICH [?]
FT   MYHIT       220    274        prf:PRO_RICH [!]
FT   MYHIT        11     34        prf:SER_RICH [?]
```

Figure 6-19: Output of Motif Scan (top).

- **The Match Detail Representation:** For each reported match, Motif Scan provides a *match detail* (a very original feature of this server) — a graph that shows you immediately whether your sequence has the right residues in the right places.

  To see how this works, take a look at the entry that corresponds to the B_ZIP domain, as shown in Figure 6-20.



Match details

match detail	match score	motif information

RpararprrpreetltpeeeekrrvrrernklaaakcrnrrreltdR	Status: !
	pos.: **136-182**
	raw-score = **49**
	N-score = **9.325**
	E-value = **0.01**
		prf:ARG_RICH
		Arginine-rich region profile.
		[entry]
		[graphics]

^ image ^		

EEKRRVRRERNKLAAAKCRNRRRELTDRLQAETDQLEEEKAELESEIAELQKEKERLEFVLVAH	Status: !
	pos.: **155-218**
	raw-score = **989**
	N-score = **13.185**
	E-value = **1.4e-06**
		prf:BZIP
		Basic-leucine zipper (bZIP) domain profile.
		[entry]
		[graphics]

Figure 6-20: Output of Motif Scan (bottom).

- **Bars *above* the sequence (green) indicate the most common amino acid for the domain at this position.** The higher the bar the more conserved this amino acid. When the bar is plain, your sequence contains the right amino acid at this position.
- **Bars *below* the sequence indicate positions where your sequence does NOT contain the right amino acid.** The length of the bar indicates how unlikely your residue is to be at this position.

  A long bar means that the amino acid found in your sequence should not be there.

This representation gives you an instant idea whether key residues (such as active sites) are conserved in your domain.

Drawing conclusions from the Motif Scan output

In Motif Scan, the graphic representation is a very powerful means of making sure you're drawing the right conclusions. For instance, with the graph on Figure 6-20, we can safely conclude that our sequence contains all the right residues for having a Basic-Leucine Zipper.

If a very conserved residue (high top bar) is replaced by a very unlikely residue in your sequence (high bottom bar), be cautious.

Note that Motif Scan also returns two domains that aren't detected by the CD server or InterProScan: the Proline-Rich Domain and the Arginine-Rich Domain. A simple glance at the corresponding sequences is enough to convince us that these are indeed Proline-Rich and Arginine-Rich domains (see Figure 6-20). This is a good example of the advantage of using more than one domain-analysis server: Now we know *two* more things about our protein.

Discovering New Domains in Your Proteins

Hunting for new domains is a bit of an art, mastered by only a few highly trained biologists around the world. But the good news is that everything you want to know is at hand — if you fancy giving it a try. The simplest way is to use BLAST, and turn your database search into what BLAST gurus call a PSSM (and simple folks call a domain). Read the BLAST chapter (Chapter 8) if you are not familiar with this tool, and you will find everything you need to build and use PSSMs at the following online address:

```
www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm
```

If you go this route, you can do everything online, but don't expect miracles. Domains are like diamonds, scattered here and there in the protein world. While you can expect to occasionally stumble by chance on a nice gem, it will take more than that to uncover the Crown Jewels! If you want to go down that road, you shouldn't mind running a few programs on Unix, digging for sequences in odd DNA sequence databases (like EST databases for instance, see Chapters 3 and 4), gathering your sequences with BLAST, aligning them with a Multiple Sequence Alignment program (see Chapter 9), turning your alignment into a Hidden Markov Model (HMM) with the Hmmer program (`hmmer.wustl.edu/`), and using Hmmer to search protein databases. This is what folks at Pfam do everyday.

REMEMBER

They say the Eskimos have 40 words for snow, and can describe even the tiniest differences. It's similar with biologists; you won't be surprised to hear that they have more than one word for protein domains. In the context of a biological paper, you can assume that the words *HMM, PSSM, profile, domain, MSA, weight matrix,* and *extended profile* mean roughly the same thing.

More Protein Analysis for Free over the Internet

The Internet offers an extremely large number of resources for doing sequence analysis online — and they're free; we've listed a few in Table 6-2. The following links are only a sample of the most stable sites available to you.

TIP

In general, if your work depends on one of these sites, we suggest that you choose a very stable resource. As a rule, avoid sites that run from a personal home page (`www.something.somewhere/~somebody`) as they're generally less reliable.

Table 6-2 Protein Sequence Analysis over the Internet

Name	*Site*	*Description*
ExPASy	`www.expasy.org/tools`	Proteins
Pbil	`npsa-pbil.ibcp.fr`	Proteins
PIR	`pir.georgetown.edu`	Proteins
CBS	`www.cbs.dtu.dk/services`	Proteins
Hits	`hits.isb-sib.ch/`	Proteins
InterPro	`www.ebi.ac.uk/interpro/scan.html`	Domains
CD search	`http://www.ebi.ac.uk/InterProScan/`	Domains

Part III

Becoming a Pro in Sequence Analysis

The 5th Wave By Rich Tennant

"That reminds me. I have to do some multiple sequence alignments later on."

In this part . . .

The most powerful tools used in bioinformatics rely on sequence comparison methods. In this part, we show you how you can search a database by sequence comparison using BLAST — or how you can use pairwise comparison techniques (such as Dotlet) to compare two sequences. We also show you multiple alignment programs — such as ClustalW or Tcoffee — that compare many sequences at the same time. If all these things are new to you, beware: They'll forever change the way you do biology!

Chapter 7

Similarity Searches on Sequence Databases

In This Chapter

- Discovering why similarity is important and what it can do for you
- Finding out everything you need to know about homology
- Running BLAST over the Internet to find a sequence similar to yours
- Determining whether your result means something
- Choosing the right flavor of BLAST
- Looking for remote homologues with PSI-BLAST
- Discovering new protein domains with PSI-BLAST

> *"When looking for a needle in a haystack, the optimistic wears gloves."*
>
> — The Little Book of Things to Keep in Mind

If you already have a protein or DNA sequence and you want to find other sequences that look like it, then you've come to the right place. In this chapter, we show you how to use BLAST to compare your sequence with every sequence contained in a sequence database — and keep the best results — in two clicks of a mouse. We also show you situations where you can use PSI-BLAST, a more powerful version of BLAST, to ask biological questions.

On the other hand, if you only know the name of your sequence or if you can only describe this sequence with words (*mouse serine and protease,* for instance) — rather than with its actual amino-acid sequence — then the first step is to go and look for this sequence in a database. We explain how to use names or descriptions to look for a sequence in Chapter 2.

Understanding the Importance of Similarity

Similar sequences often derive from the same ancestral sequence. This means that if your sequences are similar, they probably have the same ancestor, share the same structure, and have a similar biological function. This principle even works when the sequences come from very different organisms. For you, this means you can extrapolate something you know about a particular DNA or protein sequence to all similar DNA and protein sequences.

For example, imagine that your favorite sequence looks very much like another one that somebody has studied in another lab. Because these two sequences are similar, you can say, “If something is true for that sequence, it is probably true for mine as well!” Just imagine how much time you can save: Studying a gene in the lab takes years; searching a database for similarity takes seconds. And you’re not even cheating!

REMEMBER

When two proteins or gene sequences are very similar, biologists call them *homologues,* which is a fancy word for two proteins or gene sequences that have the same ancestor, similar functions, and similar structures. The snag comes in deciding how similar is “very” similar. If your sequences are more than 100 amino acids long (or 100 nucleotides long), the rule says you can label proteins as “homologous” if 25 percent of the amino acids are identical, for DNA you will require at least 70 percent identity to draw the same conclusion. If your values are below those stated values, then your guess is as good as any. You’ve entered that range of protein identity below 25 percent — known (fans of Rod Serling take note) as the *twilight zone* — where

- **Nothing is sure about the meaning of observed similarities.**

  For instance, some proteins whose amino acids are less than 15 percent identical have exactly the same 3-D structure — while some proteins with residues that are 20 percent identical have different structures.

- **Homology or non-homology is never granted.**

The 25-percent figure that defines the twilight zone is mostly a common-sense indicator. In reality, things are slightly more complicated. In most cases, to make sure that two sequences are true homologues, you also need to use some other information reported by the search. These bits of data include

- The *Expectation value* (E-value), which tells you how likely it is that the similarity between your sequence and a database sequence is due to chance
- The length of the segments similar between the two sequences

- The patterns of amino-acid conservation
- The number of insertions and deletions

Do not confuse homology and similarity. Homology is a binary relationship whereas similarity is a quantifiable property. In other words, two sequences are either homologous or non-homologous — while they will always have a measurable level of identity. As a consequence, you cannot say that two sequences are 40 percent homologous, just like you cannot say someone is 40 percent pregnant.

In the rest of this chapter, we show you how to interpret this information when you conduct a database search.

The Most Popular Data-Mining Tool Ever: BLAST

Thirty years ago, biologists would search a database by printing its whole content on paper, taping the printout to the office wall, writing down their own query sequence on a piece of paper, and spending a few hours manually scanning the wall and drinking coffee. With the millions of sequences now available, those days are gone. Fortunately, now you can use computers to run the most successful bioinformatics tool ever: BLAST, the *B*asic *L*ocal *A*lignment and *S*earch *T*ool, which is expressly designed to do the paper-on-the-wall scanning for you.

In this section, we show you two simple ways to do a BLAST search (by using a protein or a DNA sequence), and, most importantly, we show you how to interpret your results. At the end of the section, we also give you plenty of ideas on how to use BLAST for doing most of the things you could need (except making coffee).

BLASTing protein sequences

BLASTing protein sequences is what you want to do if you already have a protein sequence and you want to find other, similar protein sequences in a sequence database. You should be happy to know that BLAST is at its best with proteins. Two flavors of BLAST that exist and deal with proteins are

- **blastp:** Compares a protein sequence with a protein database.
- **tblastn:** Compares a protein sequence with a nucleotide database.

If you aren't sure what to do with your protein sequence, use blastp. When in doubt, always use blastp! Nevertheless, if you suspect that blastp may not be the best choice, determine what you're trying to discover and see Table 7-1.

Table 7-1 Choosing the Right BLAST Flavor for Proteins

What You Want	*The Right BLAST Flavor*
I want to find out something about the function of my protein.	Use blastp to compare your protein with other proteins contained in the databases.
I want to discover new genes encoding simple proteins.	Use tblastn to compare your protein with DNA sequences translated into their six possible reading frames (three on each strand).

In this section, we show you how to use two of the most popular blastp online services:

- The blastp server from its home, the National Center for Biotechnology Information (NCBI) in the USA
- The blastp server from the Swiss EMBnet server

These two servers have slightly different interfaces. If you know how to use them both, you're sure to feel at ease using any of the BLAST servers that we list at the end of this chapter.

TIP

Two good reasons for knowing how to use many BLAST servers are

- **The databases:** The BLAST servers don't give you access to the same databases. If you don't find the database you need on one server, you can always look for it on another server.
- **The speed:** The most popular servers are often overcrowded. When this happens, searching somewhere else is always an option.

WARNING!

It is a rarity for two different BLAST servers to return the same answer when you do the same query. The main reason is that the database versions often differ a little, although differences in the BLAST program version can have an effect as well. Results rarely change dramatically, but they do change a little. It's just one of those discrepancies that you must learn to live with.

Running the NCBI blastp

In this example, you are asking the following question: "Are there any proteins similar to the hamster nucleolin in the protein database Swiss-Prot?" Good question. Here's how you find the answer:

1. **Point your browser to the NCBI BLAST server at** `www.ncbi.nlm.nih.gov/BLAST`**.**

   The BLAST page at the NCBI appears, as shown in Figure 7-1.

2. **In the Protein section, click the** protein-protein BLAST (blastp) **link.**

   A search page similar to the one in Figure 7-2 appears.

3. **Paste your sequence in the search window.**

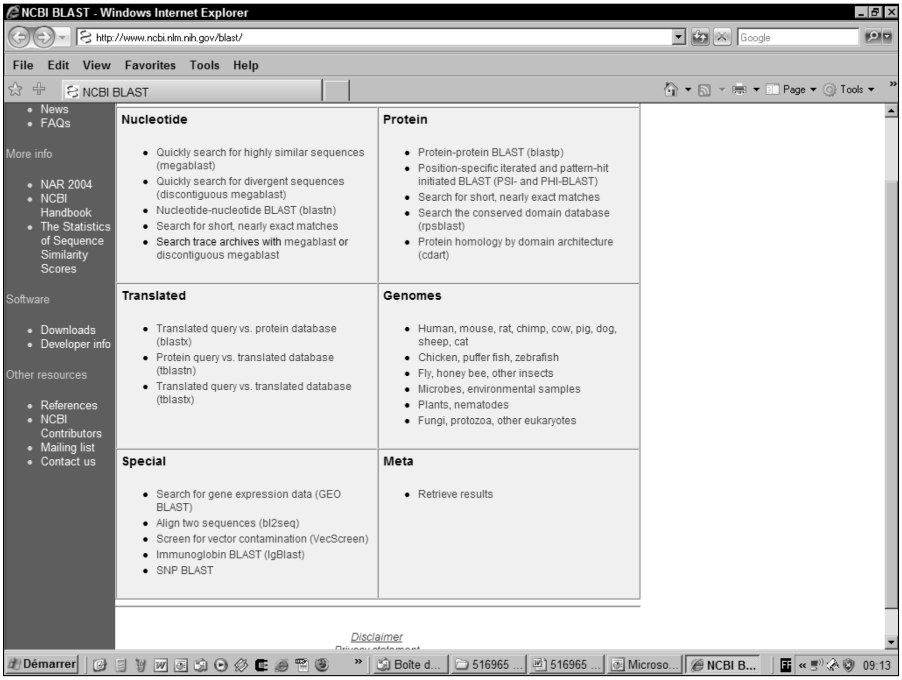
   You have two ways to pass on a sequence to the blastp server:

   - If your sequence is already in a database, you can give its ID number to blastp. For our example, we use the hamster *nucleolin,* a protein from Swiss-Prot whose accession number is P09405. To use the nucleolin, enter **P09405** in the sequence box, as shown in Figure 7-2.
   - If your sequence is not in a database, you must provide it in the FASTA format. Figure 7-3 shows you what the hamster nucleoline sequence looks like in FASTA format.

   The sequence you give to blastp is the *query sequence.*

   Sequences similar to the query that blastp returns are the *hits* or *matches.*

   The database you search is the *target database.*


NCBI BLAST - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/
Google
File Edit View Favorites Tools Help
NCBI BLAST
Page ▾ Tools ▾
Démarrer
Boîte d...
516965 ...
516965 ...
Microso...
NCBI B...
09:13

- News
- FAQs

More info

- NAR 2004
- NCBI Handbook
- The Statistics of Sequence Similarity Scores

Software

- Downloads
- Developer info

Other resources

- References
- NCBI Contributors
- Mailing list
- Contact us

Nucleotide

- Quickly search for highly similar sequences (megablast)
- Quickly search for divergent sequences (discontiguous megablast)
- Nucleotide-nucleotide BLAST (blastn)
- Search for short, nearly exact matches
- Search trace archives with megablast or discontiguous megablast

Protein

- Protein-protein BLAST (blastp)
- Position-specific iterated and pattern-hit initiated BLAST (PSI- and PHI-BLAST)
- Search for short, nearly exact matches
- Search the conserved domain database (rpsblast)
- Protein homology by domain architecture (cdart)

Translated

- Translated query vs. protein database (blastx)
- Protein query vs. translated database (tblastn)
- Translated query vs. translated database (tblastx)

Genomes

- Human, mouse, rat, chimp, cow, pig, dog, sheep, cat
- Chicken, puffer fish, zebrafish
- Fly, honey bee, other insects
- Microbes, environmental samples
- Plants, nematodes
- Fungi, protozoa, other eukaryotes

Special

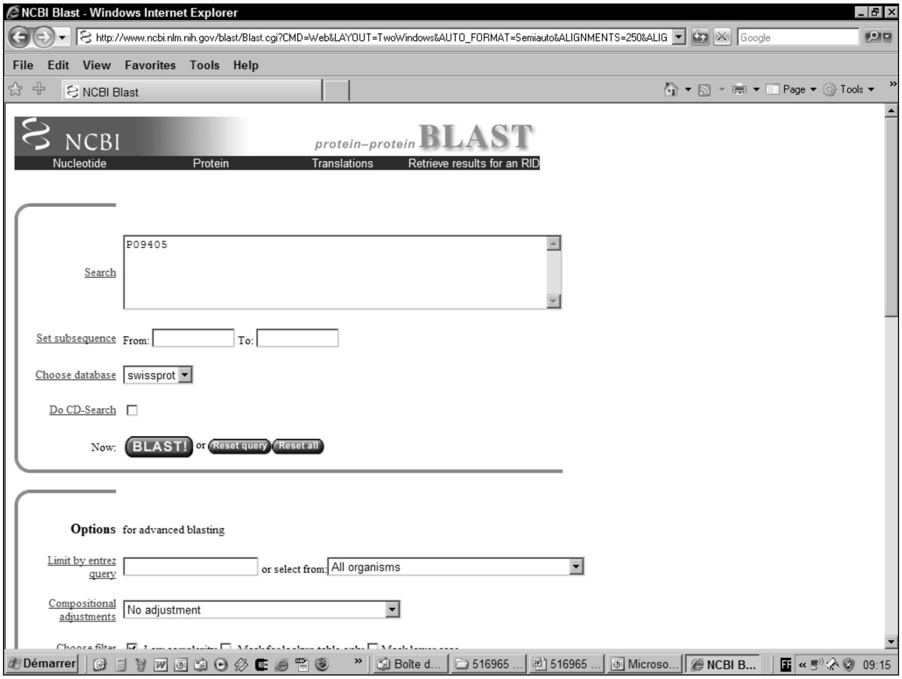
- Search for gene expression data (GEO BLAST)
- Align two sequences (bl2seq)
- Screen for vector contamination (VecScreen)
- Immunoglobin BLAST (IgBlast)
- SNP BLAST

Meta

- Retrieve results

Disclaimer

Figure 7-1: The BLAST home page at NCBI.


NCBI Blast - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIG
Google
File Edit View Favorites Tools Help
NCBI Blast
Page
Tools
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Démarrer
Boîte d...
516965 ...
516965 ...
Microso...
NCBI B...
09:15

Search: P09405

Set subsequence From: To:

Choose database: swissprot

Do CD-Search □

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query: or select from: All organisms

Compositional adjustments: No adjustment

Figure 7-2: The blastp server at NCBI.

4. **Choose Swiss-Prot from the Choose Database pull-down menu.**

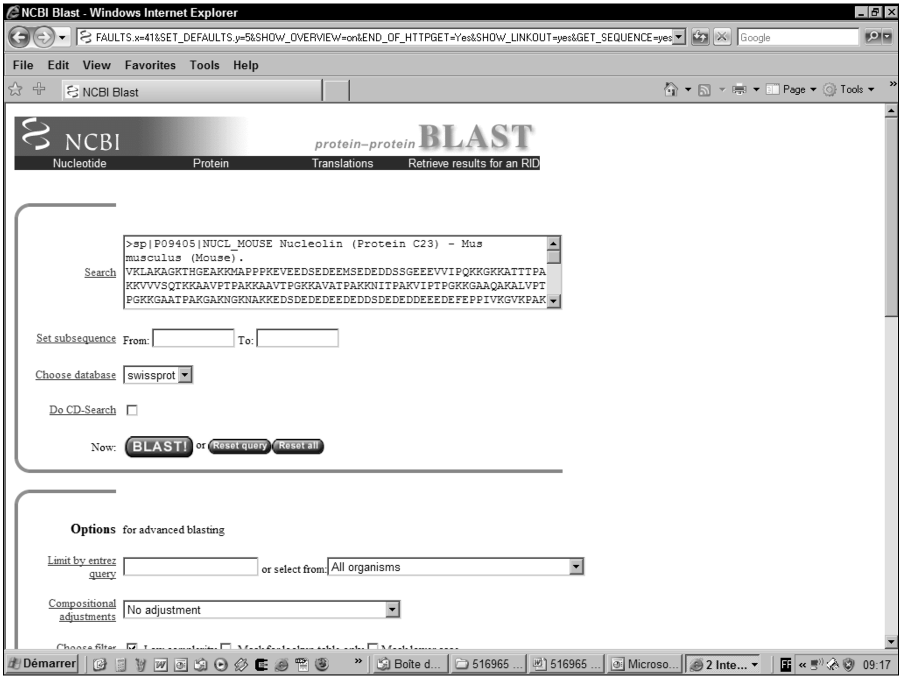
5. **Deselect the Do CD-Search box.**

   CD search is a feature for searching *C*onserved *D*omains. We recommend you deselect this box if this is one of your first BLAST searches because the output may be confusing. (We explain CD search in detail in Chapter 6.)

6. **Click the BLAST! button (and wait).**

   An intermediate page similar to Figure 7-4 appears.

   Sometimes the NCBI BLAST server gets very busy. For instance, it is notorious for scientists to dump a large number of jobs just before they go home and let them run at night. This explains why BLAST can be uncomfortable to use in the evening. Fortunately, NCBI is not the only BLAST server around; you can always try your luck on another server. (See Table 7-6, at the end of this chapter, for a listing of BLAST servers around the world.)


NCBI Blast - Windows Internet Explorer
FAULTS.x=41&SET_DEFAULTS.y=5&SHOW_OVERVIEW=on&END_OF_HTTPGET=Yes&SHOW_LINKOUT=yes&GET_SEQUENCE=yes
Google
File Edit View Favorites Tools Help
NCBI Blast
Page Tools
NCBI
protein–protein BLAST
Nucleotide Protein Translations Retrieve results for an RID
BLAST!
Reset query
Reset all
Démarrer
Boîte d...
516965 ...
516965 ...
Microso...
2 Inte...
09:17

Search

```
>sp|P09405|NUCL_MOUSE Nucleolin (Protein C23) - Mus
musculus (Mouse).
VKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTPA
KKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVPT
PGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPAK
```

Set subsequence From: To:

Choose database swissprot

Do CD-Search □

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query or select from: All organisms

Compositional adjustments No adjustment

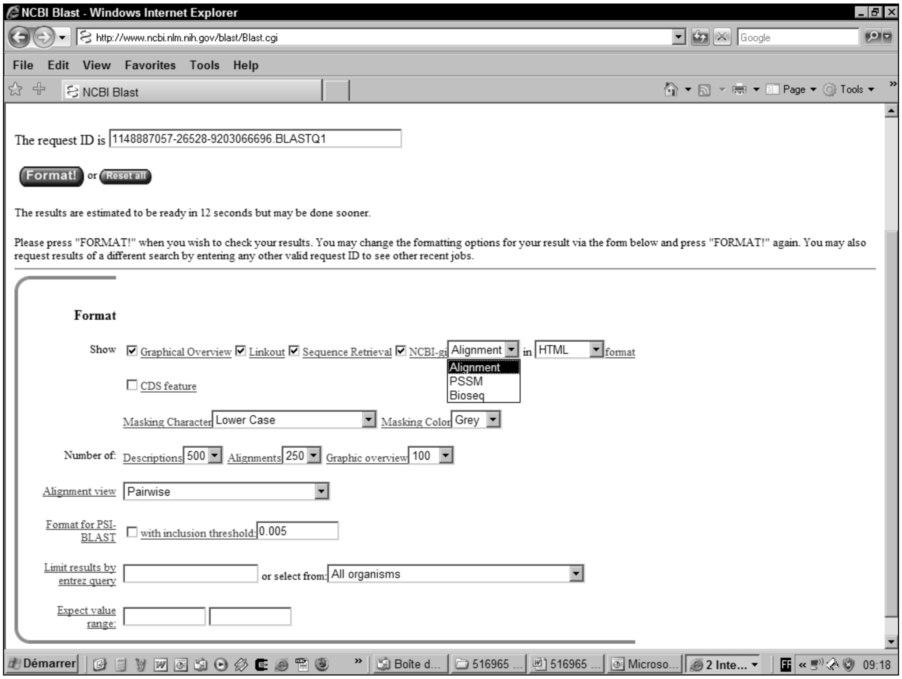
Figure 7-3: Providing the NCBI server with a sequence in FASTA format.

If the server nearest you is too slow, you can take advantage of the world's various time zones. The following table indicates the part of the world that's sleeping while you do your work in the morning or in the afternoon.

Your Location	***Morning***	***Afternoon***
USA	Japan	Europe
Europe	USA	Japan
Japan	Europe	USA

7. **Click the Format! button on the intermediate page and wait for the results.**

   When you click the Format! button, a new browser window opens. As soon as the search is complete, BLAST displays your results in this new window. Unfortunately, the search may take more time than the form indicates.


NCBI Blast - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google
File Edit View Favorites Tools Help
NCBI Blast
Page
Tools
Démarrer
Boîte d...
516965 ...
516965 ...
Microso...
2 Inte...
09:18

The request ID is 1148887057-26528-9203066696.BLASTQ1

Format! or Reset all

The results are estimated to be ready in 12 seconds but may be done sooner.

Please press "FORMAT!" when you wish to check your results. You may change the formatting options for your result via the form below and press "FORMAT!" again. You may also request results of a different search by entering any other valid request ID to see other recent jobs.

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment (Alignment / PSSM / Bioseq) in HTML format

☐ CDS feature

Masking Character Lower Case Masking Color Grey

Number of: Descriptions 500 Alignments 250 Graphic overview 100

Alignment view Pairwise

Format for PSI-BLAST ☐ with inclusion threshold: 0.005

Limit results by entrez query ____ or select from: All organisms

Expect value range: ____ ____

Figure 7-4: The BLAST intermediate result page.

A typical search against NR or Swiss-Prot takes a few minutes. If BLAST tells you that the completion of your search may take more than 20 minutes, you're probably better off trying your luck some other time — or on another BLAST server.

DO NOT press any button while you wait!

If you get no reply, DO NOT resubmit the same query several times in a row — it will only make things worse for everybody (including you)!

8. **Save your results.**

   Do not try to print the results; the alignment section may be HUGE. If you want to keep a trace of this search, save it in a file by using the File⇨Save As option of your browser.

   Unfortunately, saving a complete NCBI BLAST page isn't easy. Saved pages usually can't redisplay their active graphics when you reload them in your browser. Here are two (still rather unsatisfactory) solutions:

   - **Save the graphic display:** Pass the mouse over the graphic display, right-click, and then choose Save Picture As from the context menu that appears, as shown in Figure 7-5. The saved image is in GIF format. Although you can then include the display in electronic documents, be aware that all active hyperlinks are now lost.

- **Print your output to a PDF file:** If you have a PDF printer installed, you can print your output to a PDF file. Depending on the PDF printer you use, hyperlinks may be lost.

Running the EMBnet blastp

The BLAST server running on the EMBnet-ExPASy server is similar to the one running at NCBI, but it uses a different interface — very similar to the one used by many BLAST servers around the world. If you know how to use the EMBnet blastp, you know how to use almost any blastp server in the world.

The main difference between the NCBI blastp interface and the EMBnet interface (shown in Figure 7-6) is that the EMBnet blastp interface gives you many more choices. The downside is that you are the one who has to make sure that the various selected parameters are compatible. In the following steps, we show you how to avoid any confusion.

1. **Point your browser to** `www.ch.embnet.org/software/bBLAST.html`.

   The EMBnet Basic BLAST page appears.

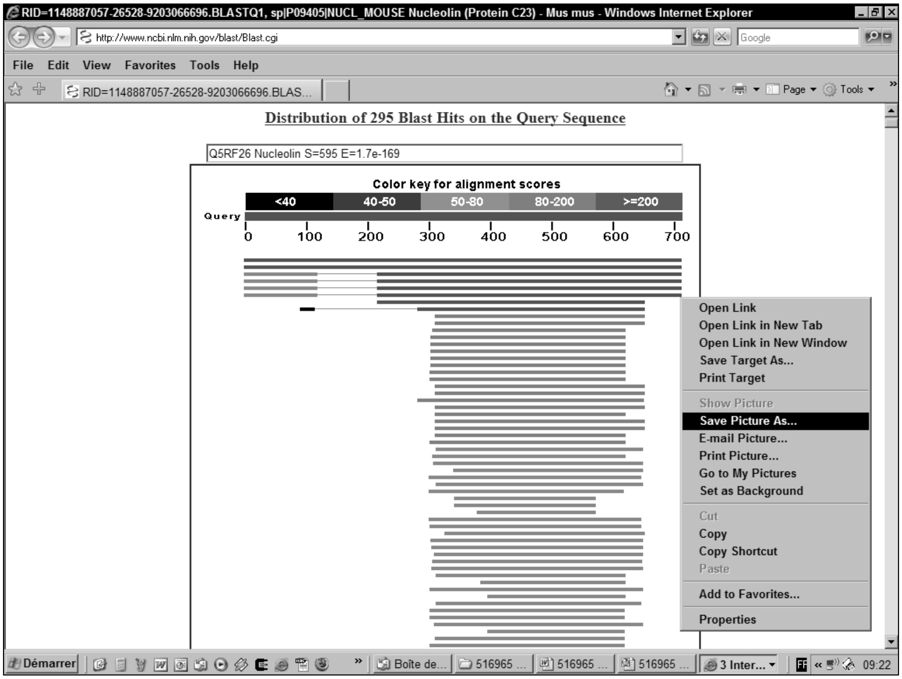


Figure 7-5: Saving the NCBI graphic display.

2. **Choose blastp from the Select the Program pull-down menu.**

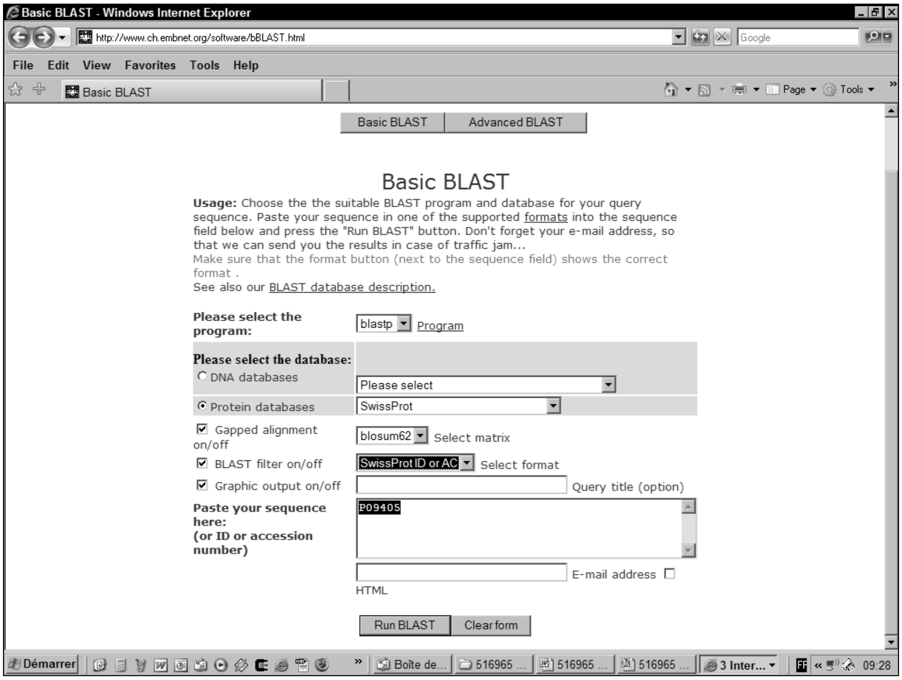
   The pull-down menu lets you select the flavor of BLAST that you want to use.

3. **After making sure that you've selected the Protein Databases radio button, select a protein database from the accompanying menu.**

   The database you select must be compatible with the BLAST program that you choose in Step 2; the EMBnet server does not check this for you. The kind of database you can search with a protein sequence as a query depends on which flavor of BLAST you selected in Step 2:

BLAST flavor	***Database type***
blastp,blastx	Protein
blastn,tblastn,tblastx	DNA

   Selecting a genome database can make it easier to analyze your results. Genome databases are those with a species name and the word *genome* in their names, such as *C.Elegans Genome.* This can be especially useful if the subject of your query belongs to a very large family. If you know which species you're interested in, choose it there. If you can't find it, use the advanced version of this server at `www.ch.embnet.org/software/aBLAST.html`.



Basic BLAST

Usage: Choose the the suitable BLAST program and database for your query sequence. Paste your sequence in one of the supported formats into the sequence field below and press the "Run BLAST" button. Don't forget your e-mail address, so that we can send you the results in case of traffic jam...
Make sure that the format button (next to the sequence field) shows the correct format .
See also our BLAST database description.

Please select the program: blastp Program

Please select the database:
○ DNA databases: Please select
● Protein databases: SwissProt

☑ Gapped alignment on/off — blosum62 Select matrix
☑ BLAST filter on/off — SwissProt ID or AC Select format
☑ Graphic output on/off — Query title (option)

Paste your sequence here: (or ID or accession number)
P09405

E-mail address □ HTML

Figure 7-6: The EMBnet BLAST interface.

If you can't find the database you need there, either, then look for it on other BLAST servers — such as the one maintained by the European Bioinformatics Institute at `www.ebi.ac.uk/blast/`.

4. **Choose Swiss-Prot ID or AC from the Select Format pull-down menu.**

   You have the choice between several formats: Plain text is a FASTA format without the first line (the one that starts with >). All the other formats refer to accession numbers, not sequences.

   It is best to keep the default in the various selectors in the boxes on the right side of the form:

   - **Gapped Alignments On/Off:** Makes it possible to force BLAST to behave like some more ancient version of this program that did not allow for gaps. To get the same behavior as the NCBI Blast, keep this box checked.
   - **BLAST Filter On/Off:** Switches off the low complexity filter. This is to prevent regions where an amino-acid is repeated many times from biasing your search. Having this on is the default, and you should keep it that way unless you have a good reason. (See the section "Controlling the sequence masking," later in this chapter.)
   - **Graphic Display On/Off:** Keep this on if you want the same graphic display as the NCBI.

5. **Enter the ID number (or the sequence itself) in the sequence box.**

   We're using **P09405** as our ID number. If you have a sequence, paste it from the Clipboard into the same window.

6. **Click the Run BLAST button.**

7. **Save your results.**

   BLAST outputs can be HUGE: Never print them out (unless you plan to rid your whole department of its stock of paper).

   When saving BLAST results, preserving the graphic display is notoriously difficult. If you want to keep this display, save it separately: Put the mouse pointer over the image, right-click, and then choose Save Picture As from the context menu that appears.

   If you have access to a PDF printer program, print your result into a PDF file; this is the best way to keep a printable electronic record.

Understanding your BLAST output

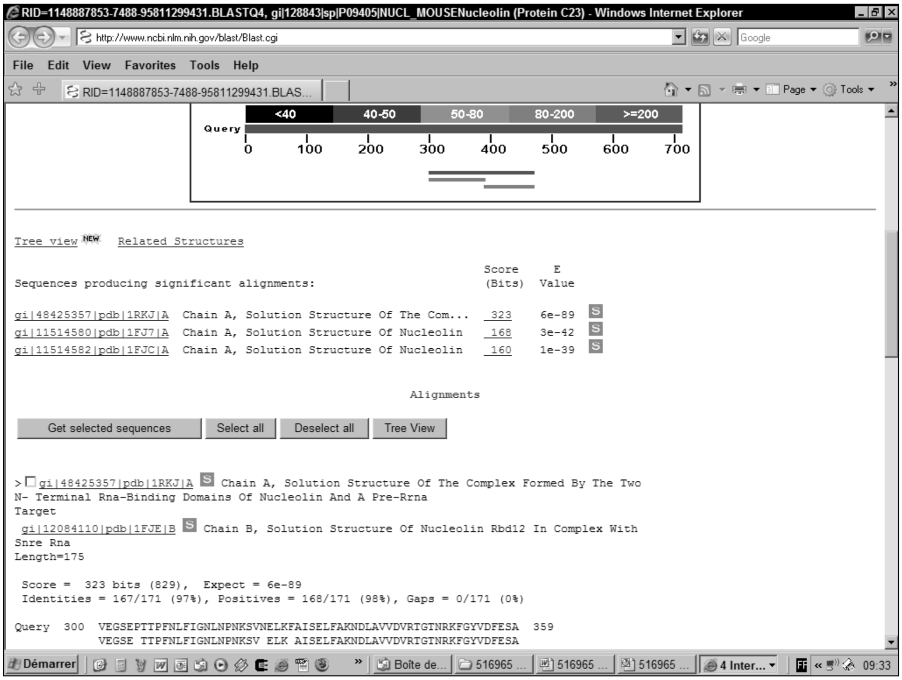
All the flavors of BLAST return a similar output. This output is rather complicated and can be confusing, even for experienced users. Figure 7-7 shows you an overview of a complete BLAST output with its four distinctive sections. If you understand everything in this output, we think you can call yourself a bioinformaticist and let others suspect you have magic powers!

An overview of the BLAST output

You'll find four sections in the output of most BLAST servers. These sections always appear in the same order, shown in Figure 7-7, and include

1. **A graphic display:** Shows you where your query is similar to other sequences. Depending on the server you use, this display can change a lot. It may also be absent on some servers.
2. **A hit list:** The name of sequences similar to your query, ranked by similarity.
3. **The alignments:** Every alignment between your query and the reported hits.
4. **The parameters:** A list of the various parameters used for the search.

Each of these elements contains a lot of information. Knowing what matters in each of these displays can be very useful to your research.


RID=1148887853-7488-95811299431.BLASTQ4, gi|128843|sp|P09405|NUCL_MOUSENucleolin (Protein C23) - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google
File Edit View Favorites Tools Help
RID=1148887853-7488-95811299431.BLAS...
Page
Tools
<40
40-50
50-80
80-200
>=200
Query
0
100
200
300
400
500
600
700
Tree view NEW
Related Structures
Get selected sequences
Select all
Deselect all
Tree View
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
4 Inter...
09:33

```
                                                                   Score     E
Sequences producing significant alignments:                        (Bits)  Value

gi|48425357|pdb|1RKJ|A  Chain A, Solution Structure Of The Com...   323    6e-89  S
gi|11514580|pdb|1FJ7|A  Chain A, Solution Structure Of Nucleolin    168    3e-42  S
gi|11514582|pdb|1FJC|A  Chain A, Solution Structure Of Nucleolin    160    1e-39  S


                                  Alignments

>□gi|48425357|pdb|1RKJ|A S Chain A, Solution Structure Of The Complex Formed By The Two
N- Terminal Rna-Binding Domains Of Nucleolin And A Pre-Rrna
Target
 gi|12084110|pdb|1FJE|B S Chain B, Solution Structure Of Nucleolin Rbd12 In Complex With
Snre Rna
Length=175

 Score =  323 bits (829),  Expect = 6e-89
 Identities = 167/171 (97%), Positives = 168/171 (98%), Gaps = 0/171 (0%)

Query  300  VEGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESA  359
            VEGSE TTPFNLFIGNLNPNKSV ELK AISELFAKNDLAVVDVRTGTNRKFGYVDFESA
```

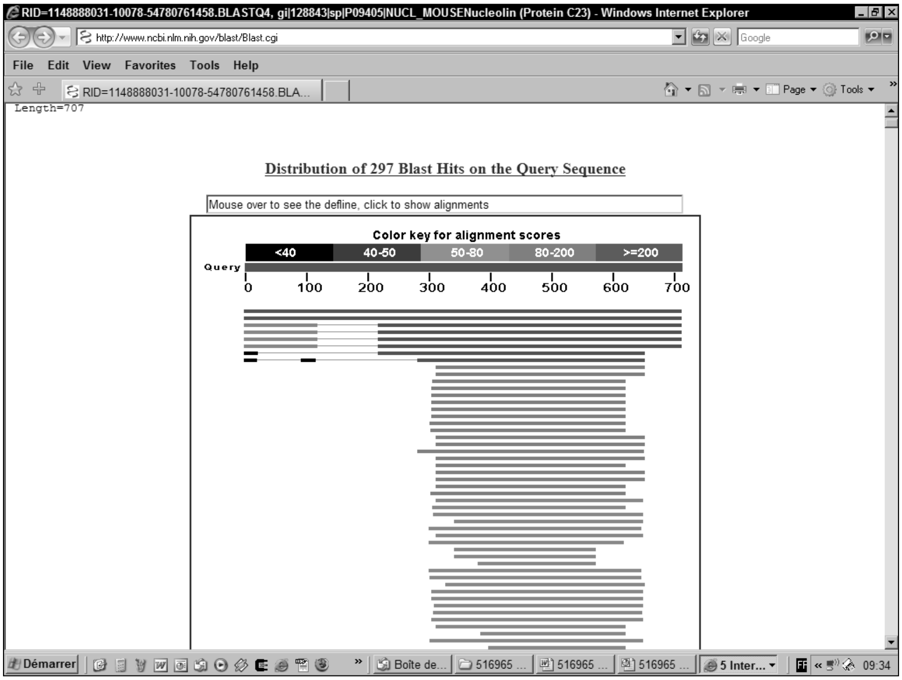
Figure 7-7: Schematic representation of the main compo nents of a BLAST output (does not include the parameters section at the bottom).

The graphic display

The display in Figure 7-8 helps you visualize your results. Your query sequence is at the top. Each bar represents the portion of another sequence that's similar to your query sequence — and the region of your sequence in which this similarity occurs. Red bars indicate the most similar sequences; pink bars indicate matches that are a bit less good; green bars indicate matches that are not impressive at all. The blue and the black bars indicate matches that have the worst scores.

Red, pink, and green matches are usually the good ones. Black hits are the bad hits: proteins that have so little in common with the query that their alignment probably means nothing from a biological point of view. (They come from the twilight zone.)

The nice thing about this display is that it can help you see that some matches do not extend over the complete length of your sequence. It is a useful tool to discover domains. For instance, here in Figure 7-8, the top hits are proteins homologous to the nucleolin. Near the bottom, the shorter hits correspond to the domain in the nucleolin that binds RNA. These hits indicate proteins that also contain this RNA binding domain but are otherwise unrelated to the nucleolin.



RID=1148888031-10078-54780761458.BLASTQ4, gi|128843|sp|P09405|NUCL_MOUSENucleolin (Protein C23) - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

Length=707

Distribution of 297 Blast Hits on the Query Sequence

Mouse over to see the defline, click to show alignments

Color key for alignment scores

Figure 7-8: The NCBI BLAST graphic display.

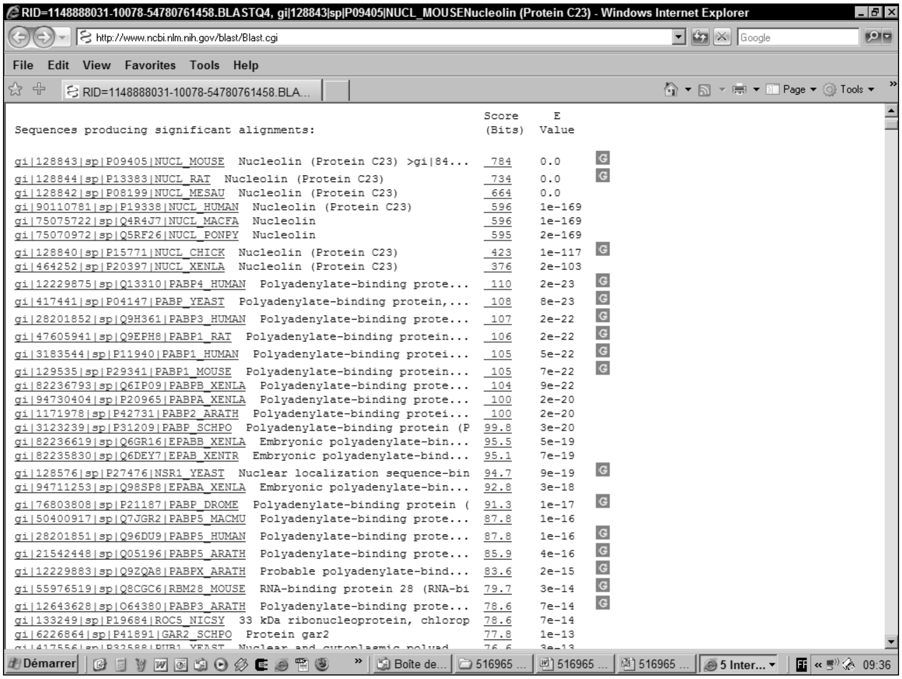
This display from the NCBI is active: If you pass the mouse pointer over a bar, the name of the corresponding sequence appears in the little window at the top; if you click this bar, the browser takes you to the corresponding alignment within this same page. (Please note that the graphical display you get from EMBnet is not active.)

When two colored lines are linked by a thin black line, as happens on the third line in Figure 7-8, it means that the same protein matches your query in two separate locations. (Please note that the thin black line is what you see with NCBI; on the EMBNet server, you'll see a dashed black line.) These matches are independent, but they could probably be joined to form a longer match.

The hit list

Most BLAST users consider the hit list (like the one shown in Figure 7-9) their favorite part of the BLAST output. It immediately tells you whether your sequence looks like something that's already in the database — and whether you can trust it as a good hit. Each line contains four important features:

- **The sequence accession number and the name:** This hyperlink takes you to the database entry that contains this sequence. In this entry, you may find very important annotated information describing the sequence. A Swiss-Prot link (|sp| in Figure 7-9) may tempt you with potentially useful information.
- **Description:** A description that comes from the annotation lets you know at a glance whether this finding could be interesting for you. Of course, you never have a *guarantee* that this annotation is correct; we recommend that you carefully check the complete annotation before getting overly excited.
- **The bit score:** A measure of the statistical significance of the alignment. The higher the bit score, the more similar the two sequences. Matches below 50 bits are very unreliable.
- **The E-value (the expectation value):** By estimating the number of times you could have expected such a good match only by chance (given the database), the E-value provides you with the most important measure of statistical significance. (See the nearby sidebar, "E-values, similarity, and homology.") The lower the E-value, the more similar the sequences and the more confidence you can have that this hit is really homologous to your query. For instance, sequences identical to your query have E-values very close to 0. Matches above 0.001 are often close to the twilight zone.
- **Genomic link:** Now that so many complete genomes are available, you probably want to know about the genomic location of potential matches. Whenever such information is available, it is indicated by a purple G. Just click the G and you'll find out everything you need to know about this gene.


RID=1148888031-10078-54780761458.BLASTQ4, gi|128843|sp|P09405|NUCL_MOUSENucleolin (Protein C23) - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google
File Edit View Favorites Tools Help
RID=1148888031-10078-54780761458.BLA...
Page
Tools
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
5 Inter...
09:36

```
                                                               Score     E
Sequences producing significant alignments:                   (Bits)  Value

gi|128843|sp|P09405|NUCL_MOUSE  Nucleolin (Protein C23) >gi|84...   784    0.0      G
gi|128844|sp|P13383|NUCL_RAT  Nucleolin (Protein C23)               734    0.0      G
gi|128842|sp|P08199|NUCL_MESAU  Nucleolin (Protein C23)             664    0.0
gi|90110781|sp|P19338|NUCL_HUMAN  Nucleolin (Protein C23)           596    1e-169
gi|75075722|sp|Q4R4J7|NUCL_MACFA  Nucleolin                         596    1e-169
gi|75070972|sp|Q5RF26|NUCL_PONPY  Nucleolin                         595    2e-169
gi|128840|sp|P15771|NUCL_CHICK  Nucleolin (Protein C23)             423    1e-117   G
gi|464252|sp|P20397|NUCL_XENLA  Nucleolin (Protein C23)             376    2e-103
gi|12229875|sp|Q13310|PABP4_HUMAN  Polyadenylate-binding prote...   110    2e-23    G
gi|417441|sp|P04147|PABP_YEAST  Polyadenylate-binding protein,...   108    8e-23    G
gi|28201852|sp|Q9H361|PABP3_HUMAN  Polyadenylate-binding prote...   107    2e-22    G
gi|47605941|sp|Q9EPH8|PABP1_RAT  Polyadenylate-binding protein...   106    2e-22    G
gi|3183544|sp|P11940|PABP1_HUMAN  Polyadenylate-binding protei...   105    5e-22    G
gi|129535|sp|P29341|PABP1_MOUSE  Polyadenylate-binding protein...   105    7e-22    G
gi|82236793|sp|Q6IP09|PABPB_XENLA  Polyadenylate-binding prote...   104    9e-22
gi|94730404|sp|P20965|PABPA_XENLA  Polyadenylate-binding prote...   100    2e-20
gi|1171978|sp|P42731|PABP2_ARATH  Polyadenylate-binding protei...   100    2e-20
gi|3123239|sp|P31209|PABP_SCHPO  Polyadenylate-binding protein (P  99.8    3e-20
gi|82236619|sp|Q6GR16|EPABB_XENLA  Embryonic polyadenylate-bin...  95.5    5e-19
gi|82235830|sp|Q6DEY7|EPAB_XENTR  Embryonic polyadenylate-bind...  95.1    7e-19
gi|128576|sp|P27476|NSR1_YEAST  Nuclear localization sequence-bin  94.7    9e-19    G
gi|94711253|sp|Q98SP8|EPABA_XENLA  Embryonic polyadenylate-bin...  92.8    3e-18
gi|76803808|sp|P21187|PABP_DROME  Polyadenylate-binding protein (  91.3    1e-17    G
gi|50400917|sp|Q7JGR2|PABP5_MACMU  Polyadenylate-binding prote...  87.8    1e-16
gi|28201851|sp|Q96DU9|PABP5_HUMAN  Polyadenylate-binding prote...  87.8    1e-16    G
gi|21542448|sp|Q05196|PABP5_ARATH  Polyadenylate-binding prote...  85.9    4e-16    G
gi|12229883|sp|Q9ZQA8|PABPX_ARATH  Probable polyadenylate-bind...  83.6    2e-15    G
gi|55976519|sp|Q8CGC6|RBM28_MOUSE  RNA-binding protein 28 (RNA-bi  79.7    3e-14    G
gi|12643628|sp|O64380|PABP3_ARATH  Polyadenylate-binding prote...  78.6    7e-14    G
gi|133249|sp|P19684|ROC5_NICSY  33 kDa ribonucleoprotein, chlorop  78.6    7e-14
gi|6226864|sp|P41891|GAR2_SCHPO  Protein gar2                      77.8    1e-13
gi|417556|sp|P32588|PUB1_YEAST  Nuclear and cytoplasmic polyad     76.6    3e-13
```

Figure 7-9: The NCBI BLAST hit list.

The alignments

No matter what we say about E-values and statistical significance, a number is only a number. When it comes to telling the real story, biologists only trust alignments. Biologists are convinced that alignments cannot lie, and this is mostly true — if you know how to look at them.

BLAST displays the alignments for a BLAST search just below the hit list, as shown in Figure 7-10. In every BLAST alignment, you can find the following features.

- **The name(s):** The first item is the name. Sometimes there is more than one sequence name; this happens whenever several sequences align identically to those in the query. For instance, if the database contains several sequences that are identical to the ones in the query, all these matches will be reported as a single hit in the alignment section. This may even happen with so-called Non-Redundant databases!
- **The percentage of identity:** This gives you a concrete substitute for the E-value. Remember the magic number: An identity of more than 25 percent is good news. (The *identity* is the number of identical residues divided by the number of matched residues — gaps are simply ignored.) The `Positives` field gives you a measure of the fraction of residues that are either identical or similar — represented with a `+` on the actual alignment. The `Gaps` field shows residues that were not aligned.

E-values, similarity, and homology

A high level of similarity between two sequences often indicates that the two have evolved from a common ancestor and have the same overall 3-D structure. Biologists call these sequences *homologues.* In practice, homologue sequences often have similar biochemical functions.

If you are studying a protein, the most desirable object in the universe is a very well-characterized protein sequence that's clearly homologous to your protein. People search databases in hopes of finding this special sequence.

That's all very fine, but how do you show that your two sequences are homologous? Think of homologous sequences as relatives in a family. We all know that relatives tend to look alike, but we also know that two persons with the same eye color aren't necessarily siblings. On the other hand, if they have the same type of hair, the same facial features, and so on, we can be tempted to conclude that they are true relatives. It works the same way with sequences.

How similar must sequences be in order to be considered homologous? The answer is clear: More than *25 percent* of the amino acids present for proteins — and more than *70 percent* of the nucleotides present for DNA — must be similar. Above this limit, you can be almost sure that two proteins have the same structure and the same common ancestor. Below that limit lies the twilight zone — that spooky identity range where nobody can really be sure whether the observed similarity means anything.

Warning: Be careful! The *25-percent* and *70-percent* limits only work for sequences that contain more than 100 amino acids or nucleotides. You might frequently get near-perfect identity between short segments (10 residues, for instance) of totally unrelated proteins or DNA sequences.

Everybody loves percent identities because they are so easy to spot visually. Unfortunately, they are not always a good indicator. For instance, they tell you nothing when many similar — but not identical — amino acids are aligned together. Moreover, how do you tell the difference between 60 matched residues spread over a 100-residue segment, and 120 matches spread over a 200-residue segment? The longest is probably more meaningful, but the percent identity says nothing about this.

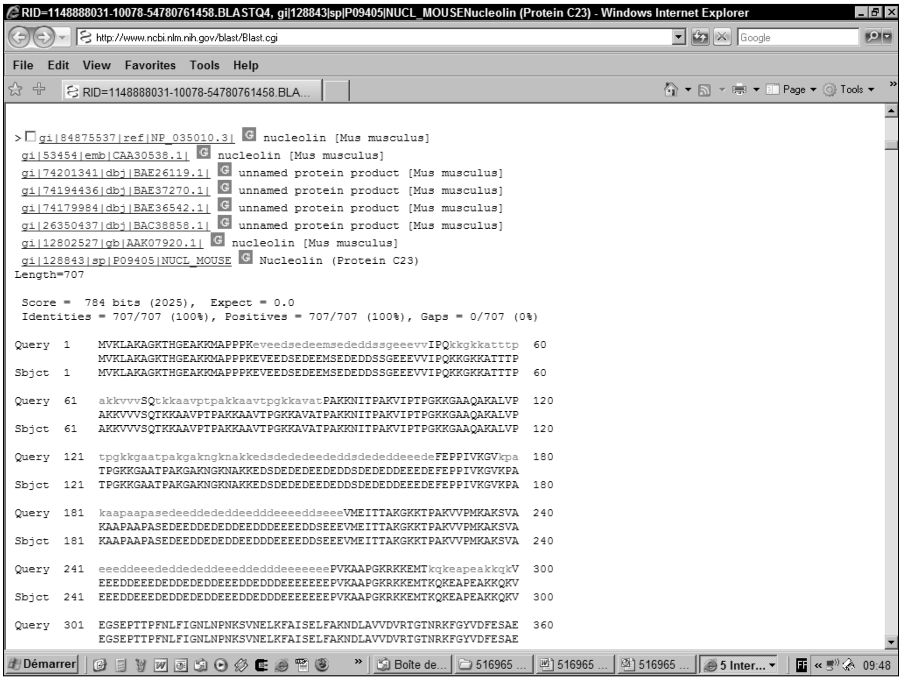
Gurus invented E-values so we'd have a criterion more objective than percentage-of-similarity. E-values (short for *expectation values*) are a powerful tool for comparing pairwise alignments with different similarities and different lengths. They also help you decide how much you can trust your conclusion on homology. With E-values, you know when you can get excited or when you should wait and see.

The E-value has a very concrete meaning: It is the number of times your database match may have occurred just by chance. We consider a match that's very unlikely to occur just by chance to be a very good match; that's why results associated with the *lowest* E-values are the best. We say they're the most *significant* because we know we can trust them enough to infer homology.

In theory, alignments associated with E-values lower than one should all be trusted. In practice, this is not true because BLAST uses an approximate formula for computing the E-values and strongly underestimates them. In the sequence world, a similarity with an E-value above 10-^{4} (0.0001) is not necessarily interesting. If you want to be certain of the homology, your E-value must be lower than 10-^{4}.

BLAST isn't the only program that uses E-values. You may come across them almost any time you compare two sequences — even if you use programs that compare sequences and domains. The principle is always the same.

- **Length:** This is the length of the alignment, which indicates how long the two segments of your sequences are that BLAST has aligned. It is the best way to see whether a hit means something. Sometimes, very short alignments can come up with high E-values and not be very meaningful.
- **The top sequence:** Your query.
- **The bottom sequence:** The "hit" — here referred to as the "subject."
- **The line in between the sequences:** Between the two sequences is a line that contains a (+) sign for similar amino acids, a letter for identical residues, or a space for mismatches.
- **The XXXXX regions:** In your query, BLAST inserts Xs automatically to mask the regions that contain many identical residues. Gurus call these regions the *low-complexity segments.* They can cause trouble in the search; the Xs tell BLAST to ignore the corresponding segments. The masking occurs only in the query sequence.
- **The numbers:** Numbers to the right side of the sequences indicate the coordinates of the match on your query sequence and on the hit sequence. BLAST makes local alignments that may contain only a portion of the query and the hit.


RID=1148888031-10078-54780761458.BLASTQ4, gi|128843|sp|P09405|NUCL_MOUSENucleolin (Protein C23) - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi
Google
File Edit View Favorites Tools Help
RID=1148888031-10078-54780761458.BLA...
Page Tools
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
5 Inter...
09:48

```
>□gi|84875537|ref|NP_035010.3| G nucleolin [Mus musculus]
 gi|53454|emb|CAA30538.1| G nucleolin [Mus musculus]
 gi|74201341|dbj|BAE26119.1| G unnamed protein product [Mus musculus]
 gi|74194436|dbj|BAE37270.1| G unnamed protein product [Mus musculus]
 gi|74179984|dbj|BAE36542.1| G unnamed protein product [Mus musculus]
 gi|26350437|dbj|BAC38858.1| G unnamed protein product [Mus musculus]
 gi|12802527|gb|AAK07920.1| G nucleolin [Mus musculus]
 gi|128843|sp|P09405|NUCL_MOUSE G Nucleolin (Protein C23)
Length=707

 Score =  784 bits (2025),  Expect = 0.0
 Identities = 707/707 (100%), Positives = 707/707 (100%), Gaps = 0/707 (0%)

Query  1    MVKLAKAGKTHGEAKKMAPPPKeveedsedeemsededdssgeeevvIPQkkgkkatttp  60
            MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP
Sbjct  1    MVKLAKAGKTHGEAKKMAPPPKEVEEDSEDEEMSEDEDDSSGEEEVVIPQKKGKKATTTP  60

Query  61   akkvvvSQtkkaavptpakkaavtpgkkavatPAKKNITPAKVIPTPGKKGAAQAKALVP  120
            AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP
Sbjct  61   AKKVVVSQTKKAAVPTPAKKAAVTPGKKAVATPAKKNITPAKVIPTPGKKGAAQAKALVP  120

Query  121  tpgkkgaatpakgakngknakkedsdededeededdsdededdeeedeFEPPIVKGVkpa  180
            TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA
Sbjct  121  TPGKKGAATPAKGAKNGKNAKKEDSDEDEDEEDEDDSDEDEDDEEEDEFEPPIVKGVKPA  180

Query  181  kaapaapasedeeddededdeedddeeeeddseeeVMEITTAKGKKTPAKVVPMKAKSVA  240
            KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA
Sbjct  181  KAAPAAPASEDEEDDEDEDDEEDDDEEEEDDSEEEVMEITTAKGKKTPAKVVPMKAKSVA  240

Query  241  eeeddeeededdededdeeeddedddeeeeeeePVKAAPGKRKKEMTkqkeapeakkqkV  300
            EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV
Sbjct  241  EEEDDEEEDEDDEDEDDEEEDDEDDDEEEEEEEPVKAAPGKRKKEMTKQKEAPEAKKQKV  300

Query  301  EGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAE  360
            EGSEPTTPFNLFIGNLNPNKSVNELKFAISELFAKNDLAVVDVRTGTNRKFGYVDFESAE
```

Figure 7-10: Pairwise alignments reported by BLAST.

A good alignment should not contain too many gaps and should have a few patches of high similarity — rather than isolated identical residues spread here and there. This criterion is especially useful if the alignment is close to the twilight zone.

Note that sometimes more than one alignment is reported within a hit. This happens when the query and the hit sequence could be aligned in several locations. On the graphic display, a thin black line (NCBI) or a dashed line (EMBnet) connects these independent alignments.

The parameters

You don't need to worry much about the meaning of the information that comes at the bottom of the result page. If you've changed the default parameters of BLAST, this part keeps track of it for you, making sure that you can reproduce this search if you need it.

Nobody can guarantee that you can reproduce a result exactly, even if you use the same BLAST server. The upgrade of any of the components in the server can modify the results of the search. Components that may be upgraded include

- The databases
- The BLAST program itself
- The default parameters of the server

In real life, you have no control over these parameters. They always end up changing with time. At least the information returned at the bottom of the search may offer a clue about what *could* be going wrong.

BLASTing DNA sequences

If you think that BLASTing a DNA sequence is like BLASTing a protein sequence, you're right — and you're not right at the same time. While it's true that the principle is the same and that in practice BLASTing DNA requires operations similar to BLASTing proteins, BLASTing DNA does not always work so well. It is faster and more accurate to BLAST proteins (blastp) rather than nucleotides.

If you know the reading frame in your sequence, you're better off translating the sequence yourself and BLASTing with a protein sequence (see the corresponding section, "BLASTing protein sequences," earlier in this chapter). If you don't know the reading frame, then you must choose one of the following BLAST programs (see Table 7-2) that deal with nucleotides.

Table 7-2 Different BLAST Programs Available for DNA Sequences

Program	*Query*	*Database*	*Usage*
blastn	DNA	DNA	Very similar DNA sequences
tblastx	TDNA	TDNA	Protein discovery and ESTs
blastx	TDNA	Protein	Analysis of the query DNA sequence

T is for *translated.* It means that you hand over a DNA sequence to BLAST, and BLAST translates this sequence into its six frames (three on one strand and three on the other one). Where there is a Stop codon, BLAST replaces it with an X. The six-frame translation means you needn't worry about the DNA strand you give to BLAST. The program tries each possible frame for you anyway.

Choosing the right flavor of BLAST for your DNA sequence

Although the list of programs in Table 7-2 with all the weird acronyms can look complicated, don't let it intimidate you — and remember that nothing ever breaks in the cyberlab! Just ask yourself the questions in Table 7-3, in the order listed, to determine which is the best program to use.

Table 7-3 Choosing the Right Flavor of BLAST for DNA

Question	*Answer*
Am I interested in non-coding DNA?	Yes: Use blastn. Never forget that blastn is only for closely related DNA sequences (more than 70 percent identical).
Do I want to discover new proteins?	Yes: Use tblastx.
Do I want to discover proteins encoded in my query DNA sequence?	Yes: Use blastx.
Am I unsure of the quality of my DNA?	Yes: Use blastx if you suspect your DNA sequence is the coding for a protein but that it may contain sequencing errors.

TIP

Blastx can correct sequencing errors for you. If you aren't sure of your protein sequence, blastx may reveal its similarity to a better-sequenced piece of DNA. If you find that your sequence contains an unexpected frameshift, don't get immediately excited. Sequencing errors cause most of the observed frame shifts — but double-check in the lab to be sure.

Making the right choices for BLASTing DNA

From a server's point of view, there's no real difference between BLASTing a protein and BLASTing a DNA sequence. The outline is exactly the same, and you can follow blindly the steps in the "BLASTing protein sequences" section, earlier in the chapter. Simply remember the following points:

- **Pick the right database:** Choose a database that's compatible with the BLAST program you want to use. Unless you use blastx, it must be a DNA sequence database. Most BLAST servers do not check to make sure that you did this properly and, consequently, let you wait forever for results that never come.
- **Restrict your search:** Database searches on DNA are slower. When possible, restrict your search to the subset of the database that you're interested in. For instance, if you're interested in the Drosophila (fruit fly), search only the Drosophila genome.
- **Shop around:** Don't hesitate to shop around to find a BLAST server that contains the database you're interested in.
- **Use filtering:** Genomic sequences are full of repetitions. Be sure to filter out at least some of them.

The BLAST way of doing things

People who do bioinformatic analysis know *there is almost nothing you can't do with BLAST!* Whatever biological question you're asking, chances are that BLAST can give you some ideas before you start using more complicated programs or spend time designing costly experiments. In Table 7-4, we give you four examples. For each of these applications, there is an exhaustive and complicated *good* solution — and then there's the BLAST way: Quick and efficient, it can be enough in most cases. (Of course, if you have too much time on your hands, you can always go for the hard way.)

Table 7-4 Asking Biological Questions with BLAST

What You Need to Do	*The BLAST Way*	*The Complicated Alternative*
Finding genes in a genome	Cut your genome sequence into little (2-to-5-kilobyte) overlapping sequences. Use blastx to BLAST each piece of genome against NR (the Non Redundant protein database). This works better if you have no introns (bacteria).	Run gene-prediction software, sequence mRNAs

What You Need to Do	*The BLAST Way*	*The Complicated Alternative*
Predicting a protein function	Use blastp to BLAST your protein sequence against Swiss-Prot. If you get a good hit (more than 25 percent identity) over the complete length of the protein, you've solved your problem and you know that your protein has the same function as the Swiss-Prot protein.	Conduct domain analysis or wet-lab experiments
Predicting a protein 3-D structure	Use blastp to BLAST your protein against PDB (the database of protein structure). If you get a good hit (an identity of more than 25 percent), you know that your protein and this good hit have a similar 3-D structure.	Conduct homology modeling, X-ray, or NMR analysis of your protein
Finding protein family members	Use blastp (or its more powerful cousin PSI-BLAST) and run it against NR (the non-redundant protein family). After you have all the members of the family, you can make a multiple sequence alignment (see Chapter 9) and draw a phylogenetic tree.	Clone new family members using PCR techniques

Controlling BLAST: Choosing the Right Parameters

As they say, *power is nothing without control.* This saying summarizes fairly well how you can get the best out of BLAST: by controlling it. In this section, we show you the main parameters in BLAST, what they do, and how you can change them to suit your needs.

When using BLAST, the #1 rule to follow is to define as clearly as possible what you want to do. Treat BLAST like an ancient oracle — to get the answer you need, your question must be crystal-clear. Tables 7-3 and 7-4 may help you fine-tune your questions so they're as clear as possible.

The default parameters that BLAST uses are quite optimal and well tested. If you don't get anything meaningful with them, don't expect any miracles if you change them. Table 7-5 lists the main reasons why you may want to change the default parameters, along with the parameters you may change. These reasons are valid for DNA and protein sequences alike.

Most of the BLAST servers use slightly different ways of specifying parameters. In this section, we have chosen to focus on the NCBI server. You will find that it is generally easy to adapt what we are saying here to your favorite server, although sometimes the parameters can feel a bit hidden. Whenever you want to customize your search, look for the advanced mode of the server you are using; that often offers more possibilities.

Table 7-5 Some Reasons to Change BLAST Default Parameters

Reason	*Parameters to Change*
The sequence you're interested in contains many identical residues; it has a biased composition.	Sequence filter (automatic masking)
BLAST doesn't report any results.	The substitution matrix or the gap penalties
Your match has a borderline E-value.	The substitution matrix or the gap penalties to check the match's robustness
BLAST reports too many matches.	The database you're searching OR filter the reported entries by keyword OR increase the number of reported matches OR increase Expect (the E-value threshold) OR reject sequences too similar to the query (those with very low E-values)

Controlling the sequence masking

When BLAST searches databases, it makes an important assumption: BLAST assumes that all your sequences are *average* sequences. For instance, if you're searching protein sequences, BLAST supposes that the average protein composition of any protein is the same as the average composition of the whole database. In practice, this assumption is too far from reality to work all the time.

Masking protein sequences

Many proteins contain patches known as low-complexity (or low-*entropy*) regions. For instance, these regions can be segments that contain many prolines or many glutamic acid residues. If BLAST aligns two proline-rich domains, this alignment gets a very good E-value because of the high number of identical amino acids it contains. Unfortunately, there is a good chance that these two proline-rich domains are not related at all. In fact, these one-amino-acid-rich domains are notorious for fooling BLAST.

To avoid this problem, BLAST filters out low-complexity regions when analyzing proteins. To do that, it replaces those regions in the sequence with Xs. If you're specifically interested in the low-complexity regions and don't want these regions filtered out of your search, you must deselect the corresponding Low Complexity check box next to Choose Filter in the Options for Advanced Blasting section of the blastp search page, as shown in Figure 7-11.

Imagine that you have just cloned and sequenced a protein. In the sequence of this protein, you find a stretch of 10 Prolines: **PPPPPPPPPP**. At this point, you're probably wondering whether this is a sequencing error or a mistake of some kind. You would feel much more confident if you had a valid protein sequence that contains the same stretch of amino acids.

To find this protein, write your own sequence **PPPPPPPPPPPPPPPPPP** and give it to blastp as a query. Do not forget to deselect the Low Complexity check box.

Some domains are very common in protein sequences, such as Zn Fingers in tandems or Fibronectines domains. If your protein contains this type of domain, BLAST reports many matches with proteins that contain the same domains but are otherwise unrelated. To make your search more interesting, filtering out these domains is a good idea. Here's how:

1. **Use CD search, InterProScan, or Pfscan to find domains in your protein.**

   See Chapter 6 for more information about using these searches.

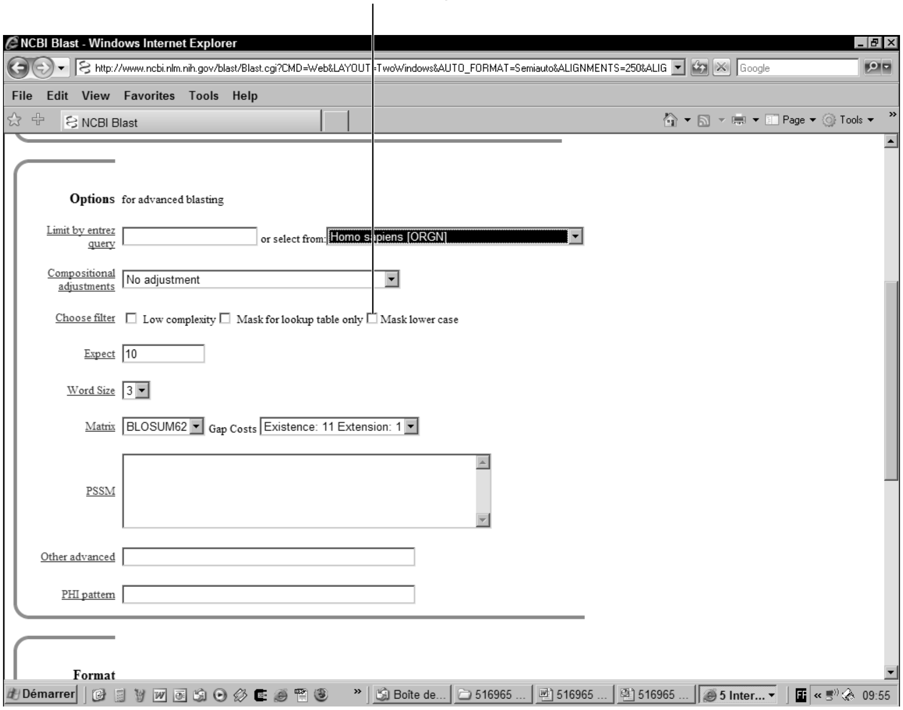
2. **Read the domain documentation to find out how widespread the domains are that you found.**

3. **Replace the sequences of less-informative domains with Xs (or rewrite them in lowercase) and then select the Mask Lower check box next to Choose Filter in the BLAST form. (See Figure 7-11.)**

4. **Run a standard blastp as shown at the beginning of this chapter.**

   When your results are returned, interpret them as shown in the "Understanding your BLAST output" section, earlier in this chapter.

Mask Lower Case option


NCBI Blast - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIG
Google
File Edit View Favorites Tools Help
NCBI Blast
Page
Tools
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
5 Inter...
FF
09:55

Options for advanced blasting

Limit by entrez query: [ ] or select from: Homo sapiens [ORGN]

Compositional adjustments: No adjustment

Choose filter: □ Low complexity □ Mask for lookup table only □ Mask lower case

Expect: 10

Word Size: 3

Matrix: BLOSUM62 Gap Costs: Existence: 11 Extension: 1

PSSM:

Other advanced:

PHI pattern:

Format

Figure 7-11: The advanced parameters of blastp at NCBI.

Masking DNA sequences

If you think it's a problem keeping your protein-sequence results significant, things get even worse when you're dealing with DNA sequences. DNA is full of sequences that repeat many times, and yet such similarities don't actually mean very much when you come right down to it. Most genomes contain many such repeats — especially the human genome, which is 60 percent repeats! If you want to avoid having to wade through that many repeats, select the Human Repeats check box that appears in the blastn page, as shown in Figure 7-12. Selecting this box causes the human repeats to be filtered out from your sequences.

TIP

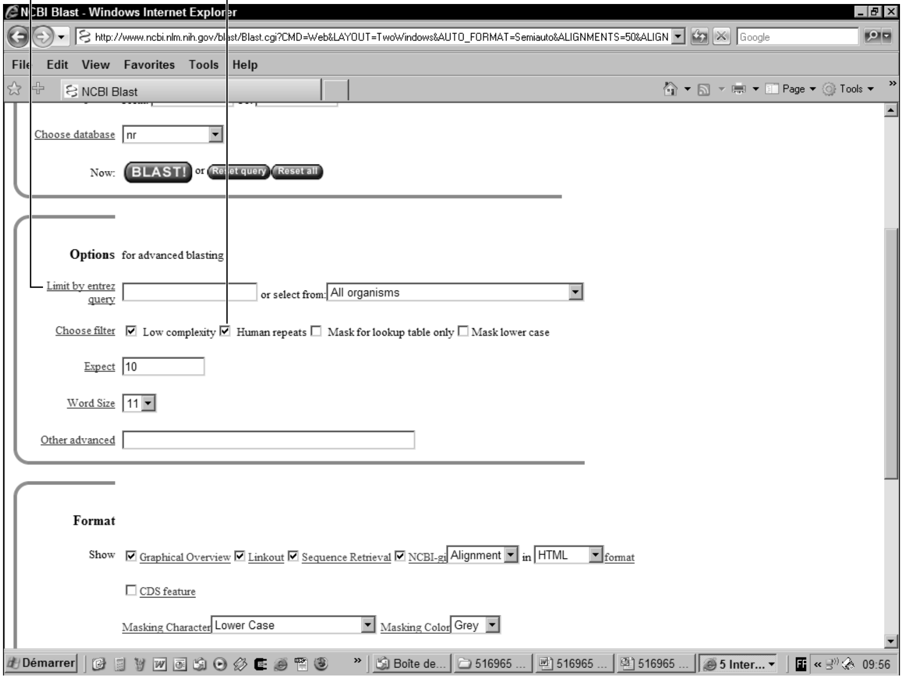
Large-scale genome sequencing is not always as smooth as the gurus want us to believe. For instance, in many cases, the final genome ends up containing bits and pieces of the organism used to clone it (which is why, when you're looking at a human genome, you may find bits of yeast and *E. coli*). This simply means that before you get very excited about a gene you've just found in the human genome (or in any genome), you may want to make sure that this gene is not a contamination brought in by the organism used for the cloning. Imagine the embarrassment if you were to discover that the cancer-susceptibility gene you've been studying for three years is in fact an *E. coli* protein!

Changing the BLAST alignment parameters

Figure 7-11 shows you all the parameters you can change on the NCBI BLAST server. These include the Expect value, the Word Size, and the means of filtering. Among these parameters, two important ones have to do with the way BLAST makes the alignments; these are the *gap penalties* (named *gap cost* on the NCBI BLAST server page) and the *substitution matrix (*named simply *matrix* on the NCBI BLAST server page). If you change either of these parameters, BLAST may report different hits. Hits that are likeliest to change from one run to the next are the hits that were borderline (those with an E-value higher than 0.0001).

Limit by Entrez Query option

Human Repeats option



NCBI Blast - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=50&ALIGN
Google
File Edit View Favorites Tools Help
NCBI Blast
Page Tools
BLAST!
Reset query
Reset all
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
5 Inter...
09:56

Choose database nr

Now: BLAST! or Reset query Reset all

Options for advanced blasting

Limit by entrez query ______ or select from: All organisms

Choose filter ☑ Low complexity ☑ Human repeats □ Mask for lookup table only □ Mask lower case

Expect 10

Word Size 11

Other advanced

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

□ CDS feature

Masking Character Lower Case Masking Color Grey

Figure 7-12: Filtering possibilities for a DNA sequence while using blastn.

Word size is BLAST's main secret recipe. In BLAST, a *word* is the minimum size of the sequence elements BLAST tries to match between your sequence and the sequences in the database. For instance, if the word size is set to 6, BLAST will consider all the possible sub-sequences of size 6 contained in your query, and it will identify in the database all the sequences containing words of size 6 similar to those in your sequence. Long words make BLAST faster and less sensitive; short words do the opposite.

The best reason to play with these parameters is to check the robustness of a borderline hit. If this match does not go away when you change the substitution matrix or the gap penalties, then it has better chances of being a biologically meaningful match.

Controlling the BLAST output

If the subject of your query belongs to a large protein family, the BLAST output may give you trouble because the databases contain too many sequences nearly identical to yours.

Sometimes this wealth of homologous sequences can prevent you from seeing a homologous sequence that's less closely related but still associated with experimental information. If this sort of thing happens to you, refer to Table 7-5 — where the last entry gives you five solutions for correcting this problem. In the following sections, we show you how to implement those solutions.

Choosing the right database

Choose a database that is suited to your needs. If BLAST reports too many hits, you may find a way around this quandary by searching Swiss-Prot rather than NR (Swiss-Prot is 100 times smaller than NR), by searching only one genome, or by only searching the PDB if you are mostly interested in structural similarities.

Limit by Entrez query

Use the Limit Results by Entrez Query box (Refer to Figure 7-12) or its associated pull-down menu if you get too many hits when you're interested in only one type of organism.

You can use this box to combine keywords with the two Boolean operators AND and NOT. For instance, if you want BLAST to report proteases only — and to ignore proteases from the HIV virus — type **"protease NOT hiv1[Organism]"** (don't forget the quotation marks). If you want to find more about this, see the online documentation on Entrez queries at

```
www.ncbi.nlm.nih.gov/entrez/query/static/help/helpdoc.html
          #Writing_Advanced_Search_Statements
```

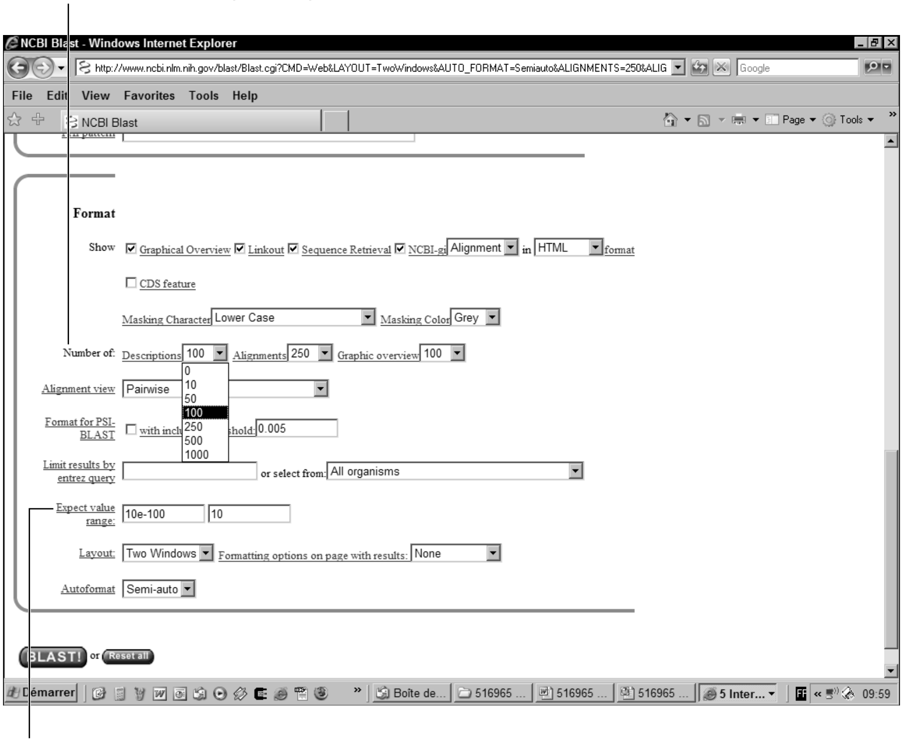
Expect

This is the cutoff point for reporting hits. BLAST doesn't report hits with an E-value higher than what you indicate in the corresponding box. A low cutoff value for this parameter forces BLAST to report only good hits.

You can filter on both side of the range and prevent the report of semi-identical sequences. For instance, in Figure 7-13, we requested that the best sequences (E-value lower than 10e–100) not be shown.

Be aware that if your search yields more than 100 hits, you also need to increase the number of reported Descriptions in the Format section, as shown in Figure 7-13.

Number of Descriptions option



NCBI Blast - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIG

Format

Show ☑ Graphical Overview ☑ Linkout ☑ Sequence Retrieval ☑ NCBI-gi Alignment in HTML format

□ CDS feature

Masking Character: Lower Case Masking Color: Grey

Number of: Descriptions 100 (0, 10, 50, 100, 250, 500, 1000) Alignments 250 Graphic overview 100

Alignment view: Pairwise

Format for PSI-BLAST: □ with incl… threshold: 0.005

Limit results by entrez query: … or select from: All organisms

Expect value range: 10e-100 10

Layout: Two Windows Formatting options on page with results: None

Autoformat: Semi-auto

Expect Value Range option

Figure 7-13: Reformatting options of BLAST.

Making BLAST Iterative with PSI-BLAST

Sometimes BLAST just isn't enough. Imagine, for instance, that you want to catch all the members of a very large protein family, starting with the one sequence that you have. When running BLAST, you catch only the most closely related sequences. How about finding the other distant cousins, the ones that your query sequence can't recognize?

PSI-BLAST does just that; first it looks for sequences that are closely related to yours — and then, gradually and carefully, it extends the circle of friends to include sequences that didn't look so interesting at first but happen to be related in the end.

Each time PSI-BLAST makes a new attempt to recruit new members of the family, it does so by using the most common properties among the members already recruited. Each new round is an *iteration.*

PSI-BLASTing protein sequences

In the following example, we use PSI-BLAST to ask whether leghemoglobin, a protein that sequesters oxygen in plants, is related to human hemoglobin.

1. **Point your browser to the main NCBI BLAST page at**

   ```
   www.ncbi.nlm.nih.gov/BLAST
   ```

2. **Click the PSI- and PHI-BLAST link (the second link under** `protein BLAST`**).**

   The PSI-BLAST page appears, as shown in Figure 7-14.

3. **Enter your sequence accession number or paste the complete sequence in the window.**

   The procedure to give a sequence to PSI-BLAST is exactly the same as the procedure blastp uses. (Refer to Figures 7-2 and 7-3.)

   - If your sequence is already in a database, you can type in its ID. For this example, we use the human hemoglobin, a protein from Swiss-Prot whose accession number is P01922.
   - If your sequence isn't in a database, you must provide it in the FASTA format.

4. **Choose PDB from the Choose Database pull-down menu (refer to Figure 7-14).**

   PDB is the Protein Database, a database of proteins with a known 3-D structure. We chose it here because it is a small database that lends itself well to this example.

The aim of the search is to hit a PDB sequence annotated as a leghemoglobin, and therefore convince ourselves that hemoglobin and leghemoglobin are structurally related.

TIP

If the first PSI-BLAST run finds no homologue to your sequence, you can use NR (the complete Non-Redundant protein-sequence database) instead of PDB. NR may contain intermediates between your sequence and another PDB sequence that PSI-BLAST may catch after a few iterations.

5. **Click the BLAST! button.**

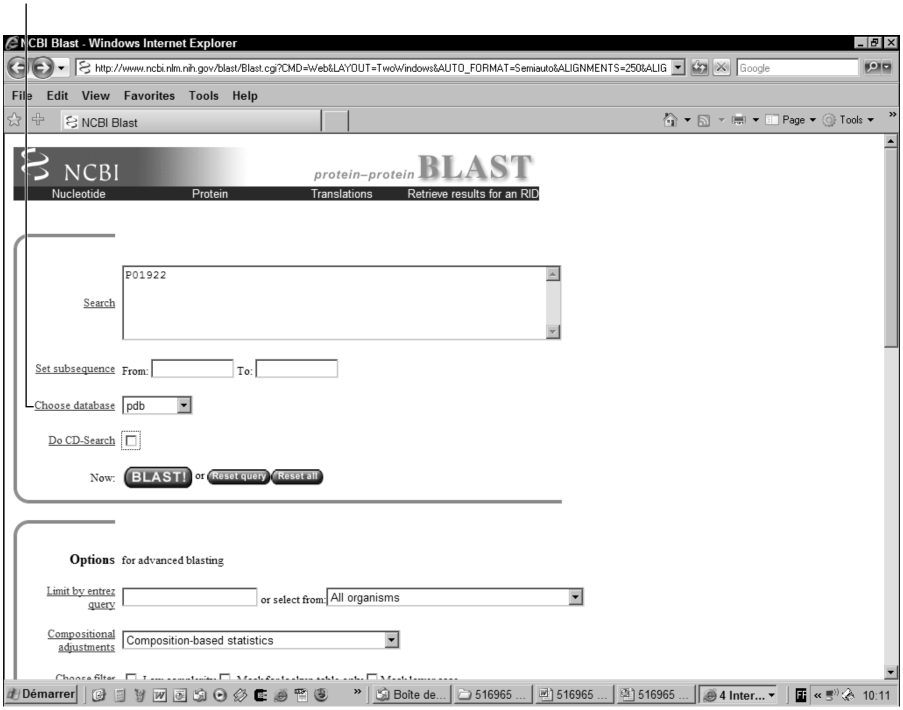
   The time it takes can be longer than what it says on-screen. Be patient!

   An intermediate page (titled Reformatting BLAST) appears, containing a Format! button.

6. **Click this Format! button.**

   A new page appears in a new window titled Results of BLAST. This is where your results are displayed when ready.

Choose Database menu



Search: P01922

Set subsequence From: To:

Choose database: pdb

Do CD-Search □

Options for advanced blasting

Limit by entrez query: or select from: All organisms

Compositional adjustments: Composition-based statistics

NCBI Blast - Windows Internet Explorer
http://www.ncbi.nlm.nih.gov/blast/Blast.cgi?CMD=Web&LAYOUT=TwoWindows&AUTO_FORMAT=Semiauto&ALIGNMENTS=250&ALIG
Google
File Edit View Favorites Tools Help
NCBI Blast
Page Tools
NCBI
protein–protein BLAST
Nucleotide
Protein
Translations
Retrieve results for an RID
Now:
BLAST!
or
Reset query
Reset all
Choose filter
Démarrer
Boîte de...
516965 ...
516965 ...
516965 ...
4 Inter...
10:11

Figure 7-14: Giving a sequence to PSI-BLAST.

Be sure to keep the Reformatting BLAST window open: You're going to need it.

7. **Inspect the results.**

   There are many very similar sequences and only a few distantly related. Getting excited here is difficult; we have many hemoglobin proteins and a few myoglobins, as we might have expected.

   Note in Figure 7-15 that the hit list in PSI-BLAST is different from the one we see in BLAST proper. It contains three new features associated with each sequence:

   - **A check box:** If this box is selected, PSI-BLAST is going to use the corresponding sequence to derive the position-specific matrix for the next iteration of PSI-BLAST.
   - **The New symbol:** It shows you sequences that PSI-BLAST reports as hits for the first time.
   - **The green pill:** It shows you sequences that PSI-BLAST has already used to obtain the current result.

   If the description makes you think one of the sequences is not related to your initial query, don't hesitate to uncheck it. That way, the unrelated sequence won't be used in the next iteration.

8. **Click the Run PSI-BLAST Iteration 2 button (near the top of the page).**

   The Reformatting BLAST window pops up.

9. **Click the Format! button on the Reformatting BLAST window.**

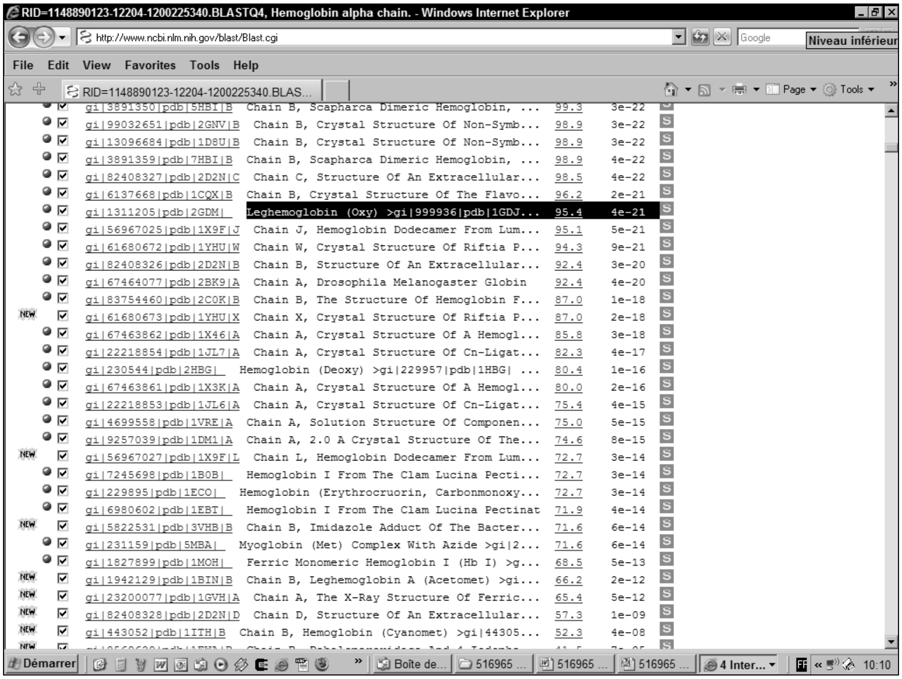
   The results appear in the Result of BLAST window.

10. **Continue repeating Steps 9 and 10.**

    Leghemoglobin appears at the second iteration, but it doesn't have a very good E-value. However, after the third iteration (Figure 7-15), its E-value improves noticeably, suggesting that it is a genuine relative of the hemoglobin family.

Avoiding mistakes when running PSI-BLAST

In the leghemoglobin example (in the preceding section), we keep going from one iteration to the next because the annotated sequences that every iteration adds are obviously hemoglobin-related. That's a sign we're going in the right direction.



RID=1148890123-12204-1200225340.BLASTQ4, Hemoglobin alpha chain. - Windows Internet Explorer

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi

```
●   ☑ gi|3891350|pdb|5HBI|B  Chain B, Scapharca Dimeric Hemoglobin, ...  99.3  3e-22
●   ☑ gi|99032651|pdb|2GNV|B  Chain B, Crystal Structure Of Non-Symb...  98.9  3e-22
●   ☑ gi|13096684|pdb|1D8U|B  Chain B, Crystal Structure Of Non-Symb...  98.9  3e-22
●   ☑ gi|3891359|pdb|7HBI|B  Chain B, Scapharca Dimeric Hemoglobin, ...  98.9  4e-22
●   ☑ gi|82408327|pdb|2D2N|C  Chain C, Structure Of An Extracellular...  98.5  4e-22
●   ☑ gi|6137668|pdb|1CQX|B  Chain B, Crystal Structure Of The Flavo...  96.2  2e-21
●   ☑ gi|1311205|pdb|2GDM|   Leghemoglobin (Oxy) >gi|999936|pdb|1GDJ...  95.4  4e-21
●   ☑ gi|56967025|pdb|1X9F|J  Chain J, Hemoglobin Dodecamer From Lum...  95.1  5e-21
●   ☑ gi|61680672|pdb|1YHU|W  Chain W, Crystal Structure Of Riftia P...  94.3  9e-21
●   ☑ gi|82408326|pdb|2D2N|B  Chain B, Structure Of An Extracellular...  92.4  3e-20
●   ☑ gi|67464077|pdb|2BK9|A  Chain A, Drosophila Melanogaster Globin    92.4  4e-20
●   ☑ gi|83754460|pdb|2C0K|B  Chain B, The Structure Of Hemoglobin F...  87.0  1e-18
NEW ☑ gi|61680673|pdb|1YHU|X  Chain X, Crystal Structure Of Riftia P...  87.0  2e-18
●   ☑ gi|67463862|pdb|1X46|A  Chain A, Crystal Structure Of A Hemogl...  85.8  3e-18
●   ☑ gi|22218854|pdb|1JL7|A  Chain A, Crystal Structure Of Cn-Ligat...  82.3  4e-17
●   ☑ gi|230544|pdb|2HBG|   Hemoglobin (Deoxy) >gi|229957|pdb|1HBG| ...  80.4  1e-16
●   ☑ gi|67463861|pdb|1X3K|A  Chain A, Crystal Structure Of A Hemogl...  80.0  2e-16
●   ☑ gi|22218853|pdb|1JL6|A  Chain A, Crystal Structure Of Cn-Ligat...  75.4  4e-15
●   ☑ gi|4699558|pdb|1VRE|A  Chain A, Solution Structure Of Componen...  75.0  5e-15
●   ☑ gi|9257039|pdb|1DM1|A  Chain A, 2.0 A Crystal Structure Of The...  74.6  8e-15
NEW ☑ gi|56967027|pdb|1X9F|L  Chain L, Hemoglobin Dodecamer From Lum...  72.7  3e-14
●   ☑ gi|7245698|pdb|1B0B|   Hemoglobin I From The Clam Lucina Pecti...  72.7  3e-14
●   ☑ gi|229895|pdb|1ECO|   Hemoglobin (Erythrocruorin, Carbonmonoxy...  72.7  3e-14
●   ☑ gi|6980602|pdb|1EBT|   Hemoglobin I From The Clam Lucina Pectinat  71.9  4e-14
NEW ☑ gi|5822531|pdb|3VHB|B  Chain B, Imidazole Adduct Of The Bacter...  71.6  6e-14
●   ☑ gi|231159|pdb|5MBA|   Myoglobin (Met) Complex With Azide >gi|2...  71.6  6e-14
●   ☑ gi|1827899|pdb|1MOH|   Ferric Monomeric Hemoglobin I (Hb I) >g...  68.5  5e-13
NEW ☑ gi|1942129|pdb|1BIN|B  Chain B, Leghemoglobin A (Acetomet) >gi...  66.2  2e-12
NEW ☑ gi|23200077|pdb|1GVH|A  Chain A, The X-Ray Structure Of Ferric...  65.4  5e-12
NEW ☑ gi|82408328|pdb|2D2N|D  Chain D, Structure Of An Extracellular...  57.3  1e-09
NEW ☑ gi|443052|pdb|1ITH|B  Chain B, Hemoglobin (Cyanomet) >gi|44305...  52.3  4e-08
```

Figure 7-15: The hit list in PSI-BLAST.

On the other hand, sometimes the road is a dead end. If, after the second iteration, every new hit is an alcohol dehydrogenase, that means it's time to give up! Unfortunately, no simple rule governs when to stop or when to continue. The only source of help is the annotation. Anything strange happening means one of two things: that you must stop — or that you have found something truly interesting.

REMEMBER

Bear in mind that you also have the obvious possibility (and even the obligation!) to uncheck clearly false matches between two iterations.

Choosing the right sequences

The main difficulty in PSI-BLAST is deciding which sequences you can keep from one iteration to the next. PSI-BLAST controls that decision automatically with the E-value threshold (Inclusion Threshold in the Format section) — but you can use the check box associated with each match to alter the automatic procedure manually.

This Inclusion Threshold is a mixed blessing, for the following reasons:

- If you set this threshold too low, PSI-BLAST doesn't catch any sequence.
- If you set this threshold too high, PSI-BLAST catches unrelated sequences.

Catching unrelated sequences is really like catching fleas: It gets worse and worse. The first bad sequence you catch attracts other sequences unrelated to your query, and things rapidly get out of control.

The only good reason to lower the threshold is a total lack of results from the first iteration. Otherwise a good strategy is to keep the default settings and inspect the hits that are below the threshold. After each iteration, you can also add the hits that seem relevant (according to their annotations) and remove those that are selected but seem irrelevant. All this takes time — and it is a bit of an art — but sometimes it pays off.

Changing the query is also a means to check your results. If you use one of the matches as an initial query, you can reasonably expect that PSI-BLAST may find your original query after a few iterations.

Avoiding the confusion of multi-domain proteins

When you consider that a domain is something like a small autonomous protein that can fold on its own, it's no surprise that most proteins are multidomains. The same domain may occur in very different proteins.

If your protein is multi-domain, PSI-BLAST can run into trouble because each domain competes with the others — and they may all recruit different sequences from unrelated proteins. The result you get from such a query is usually very difficult to interpret.

When this ambiguity happens, the best strategy is to cut your protein according to its domain structure — and then treat each sequence independently. You can find where the domains are by using the NCBI CD server or the EMBnet pfscan. (See Chapter 6 for more details.)

TIP

As a rule, if you didn't find any domain — and if your protein is larger than 200 amino acids — cut it into smaller pieces, each 200 amino acids long, and analyze them one by one. The snag in this strategy is that it leaves you with the boring task of putting together all the pieces of the puzzle. But sometimes that's where the fun begins.

Discovering and using protein domains with BLAST and PSI-BLAST

In Chapter 6, we show you how to look for known protein domains using the CD server (also known as *reverse-PSI-BLAST* or *rps-blast*). Even in that scenario, you can use BLAST to discover your own domains and use them to scan protein databases. In the BLAST dialect, a domain is named a *PSSM* (*P*osition *S*pecific *S*ubstitution *M*atrix).

Whenever you run PSI-BLAST, you can output your results in the form of a PSSM. In order to do that, follow the simple steps outlined in the BLAST online documentation at

```
www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm
```

Once your PSSM is ready, all you need do is a cut and paste it into the corresponding section of the advanced BLAST parameters form. (See Figure 7-11.)

When you do provide a PSSM, BLAST uses it in place of a single query sequence — and will compare it against every sequence in the database you choose.

Similarity Searches for Free over the Internet

The BLAST online documentation is terrific — you should use it as much as you can! The most informative pages are

- `www.ncbi.nlm.nih.gov/blast/producttable.shtml`

  This one's perfect for helping you choose the right flavor of BLAST and the right database.

- `www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml`

  This one tells you everything you need to know about each parameter.

If you've tried to run the examples in this chapter on the NCBI Web server, you know that sometimes there's just too much traffic to do useful work. Fortunately, BLAST servers keep popping up all around the world. Those we show in Table 7-6 constitute only a sample of the most robust and up-to-date servers.

Table 7-6 BLAST and PSI-BLAST Servers around the World

Country/ Continent	*Program*	*URL*
USA	BLAST/PSI-BLAST	`www.ncbi.nlm.nih.org/BLAST`
Europe	BLAST	`www.expasy.ch/tools/blast/`

(continued)

Table 7-6 *(continued)*

Country/ Continent	*Program*	*URL*
Europe	BLAST	`www.ch.embnet.org/software/bBLAST.html`
Europe	BLAST	`www.ebi.ac.uk/blast`
Japan	BLAST/PSI-BLAST	`www.ddbj.nig.ac.jp/search/blast-e.html`

The NCBI-BLAST has a little sibling, known as WU-BLAST, where WU stands for Washington University. Specialists can spend hours arguing about why WU-BLAST is more sensitive and more gifted at inserting gaps than the NCBI-BLAST. The best argument here would probably be that you should try it for yourself (See Table 7-7) and join the debate . . .

Table 7-7 WU-BLAST

Address	*Description*
`blast.wustl.edu/`	The Home of WU-BLAST (no online server)
`tigrblast.tigr.org/tgi/`	WU-BLAST at TIGR, with a twist on bacterial genomes and nucleotide analysis
`www.genome.wustl.edu/tools/blast/`	blastn and tblastn, for digging into complete genomes
`www.ebi.ac.uk/blast/`	All flavors of BLAST and WU-BLAST
`brassica.bbsrc.ac.uk/BrassicaDB/blast_form.html`	Another complete WU-BLAST server

Although it's by far the most popular, BLAST isn't the only method available for searching databases. Three main alternatives to BLAST are

- **Smith and Waterman (SSEARCH):** It's slower — but arguably more accurate — than BLAST.
- **FASTA:** It's a bit slower than BLAST but allegedly more accurate when making DNA comparisons.
- **BLAT:** Use this for locating cDNA rapidly in a genome or finding close (mammalian-versus-mammalian) proteins in a genome.

In practice, the interface for using these methods is very similar to those used for BLAST. You should have no problem adapting the content of this chapter to any similarity-based database search over the Internet. In Table 7-8, we give you a sample of the most stable and up-to-date servers that run these services.

None of these programs has any BLAST-type facility for filtering low-complexity regions.

Table 7-8 Alternative Methods for Homology Searches

Country/ Continent	*Program*	*Address*
USA	FASTA	`fasta.bioch.Virginia.edu/fasta`
Europe	FASTA	`www.ebi.ac.uk/fasta33`
Europe	SSEARCH	`www.ch.embnet.org/software/GMFDF_form.html`
Japan	SSEARCH/FASTA	`www.ddbj.nig.ac.jp/search/ssearch-e.html`
USA	BLAT	`genome.ucsc.edu`

Chapter 8

Comparing Two Sequences

In This Chapter

- ▶ Making a dot plot over the Internet
- ▶ Identifying repeats in a protein sequence
- ▶ Finding out about local and global alignments
- ▶ Finding the best local alignments between your two sequences
- ▶ Getting a handle on the gap penalty and the substitution matrix

Your true value depends entirely on what you are compared with.

— Bob Wells

In this chapter, we show you some of the methods available for comparing two protein sequences — or two DNA sequences — with each other. In their own peculiar jargon, bioinformatics gurus call these analyses *pairwise comparisons* because they relate to a pair of sequences. They are extremely useful and can help you

- ✓ Convince yourself that two sequences are in fact homologous
- ✓ Find out that your sequences share a domain
- ✓ Identify the exact location of common features, such as disulfide bridges or catalytic active sites
- ✓ Compare a gene and its product

Typically, biologists use pairwise comparisons to refine database search results and carry out detailed analysis. It is common practice to use pairwise techniques after a database search in order to check whether some of the reported matches are really interesting or are just a flash in the pan.

In this chapter, we show you which method is best suited to your needs — and also show you how you can use the Internet to run a method on the sequences you're interested in.

Making Sure You Have the Right Sequences and the Right Methods

The most delicate decision you make, when doing a pairwise analysis, is the choice of two sequences to compare. Choosing two sequences is a bit like arranging a boxing match between two opponents: The idea is to get the most exciting fight.

WARNING!

Don't use pairwise comparison methods to discover a sequence that would be homologous to a sequence you already have. It takes too much time. If you want to compare your sequence with every other sequence in a database, use a database-search program such as BLAST. (See Chapter 7 for more on BLAST.)

If you've made your way through Chapter 7, you're now in a position to argue that database-search programs merely do pairwise comparisons between a query sequence and all the sequences within a database — so there's no real need to make extra pairwise comparisons, is there? But wait a minute. Although it's true that programs like BLAST search databases through pairwise comparisons, these programs are optimized for speed, not for alignment accuracy. The programs we describe here are just the opposite: They are optimized for giving the most accurate possible result, which is why you want to apply them to carefully selected sequences.

Choosing the right sequences

A good reason to make a pairwise comparison between two sequences is a strong suspicion that these sequences are *homologous* — that is, they share a common ancestor. (See Chapter 7 for a complete description.) Such sequences often have similar 3-D structures and related functions.

The best way to find a sequence that's homologous to a sequence you already have is to search a database with the BLAST program. Select your sequences according to the following (conservative) criteria:

- **DNA sequence:** At least 70 percent identity over more than 100 bases between the hit and the query, or an E-value lower than 10-^{4}. (For more on E-values, see Chapter 7.)
- **Protein sequence:** More than 25 percent identity over more than 100 amino acids between the hit and the query, or an E-value lower than 10-^{4}.

Unfortunately, these criteria are only an indication of a match, not a guarantee. If your hit has a score very close to the threshold, it may not be homologous to the query — or it may be so distantly related that aligning it correctly is difficult. This is where pairwise comparison methods come into the picture: They help you decide how meaningful a database hit really is.

Making a database search is like scanning the classified ads for a new apartment. You may find a few interesting things, but you can't be sure you've found the perfect (or perhaps just adequate) living space until you visit. The "penthouse with air conditioning and an astonishing view" may in fact be a run-down shed with gaping holes in the wall on top of a factory building.

Choosing a sequence from a database search output is all about making the right compromise between new information and certainty: You want a sequence that's similar enough to your query so you can be sure it's homologous. On the other hand, you want this sequence to be interesting from a biological point of view so the comparisons reveal something new to you.

In short, the object of your desire is an extensively annotated Swiss-Prot sequence that matches your query with an E-value much lower than 10-^{4}. In real life, we often have to compromise.

No written rule says you *must* select your sequences from a database search — doing so is simply very convenient. An alternative is to select your two sequences by using an experimental criterion: say, because they have the same function or other similar property. For this purpose, you may use the Sequence Retrieval System (SRS) that enables you to identify sequences by using keywords. (See Chapter 2 for more on this topic.)

If you have only one sequence, it is possible (and sometimes advisable) to make a pairwise comparison on this sequence alone. All you need to do is pretend that the second sequence is identical to the first one. This may seem a bit silly, but it's a good way to discover interesting features in your sequence, such as

- Repeated domains
- Regions with a small motif repeated many times (low complexity)
- Palindromes (portions of DNA that are repeated in different orientations) and potential secondary structures in RNA

Internal repeats can help elucidate the function of your protein. For instance, the discovery of a repeated region in the breast-cancer-susceptibility gene of type 1 (Brca1) made it possible to build a domain profile for this domain — and to identify the involvement of Brca1 in DNA repair mechanisms. Even by itself, this story makes a good case for not overlooking self-comparisons!

Choosing the right method

Comparing sequences is one of the most difficult tasks biologists ever came across. In fact, the problem is not entirely solved yet — which is why there are more solutions than one. Table 8-1 lists three methods and indicates the situations for which each is best suited.

Table 8-1 Different Types of Pairwise Comparisons

Method Name	*Situation*
Dot plot	General exploration of your sequence: Discovering repeats Finding long insertions/deletions Extracting portions of sequences to make a multiple alignment
Local alignments	Comparing sequences with partial homology: Making high-quality alignments Making residue-per-residue analysis
Global alignments	Comparing two sequences over their entire length: Identifying long insertions/deletions Checking the quality of your data Identifying every mutation in your sequences

It's a good strategy always to start with a *dot plot.* (We get into all the dot-plot details in the following section.) Dot plots are very powerful tools that give you an instant general picture. With a bit of experience, a simple glance at a dot plot immediately tells you what is really going on. It's an ideal tool for deciding on the next step of your analysis.

Yet, a dot plot is not enough for a fine-grained examination. Dot plots do not really produce alignments; they simply give generic indications. They don't tell unambiguously how amino acids or nucleotides correspond to one another between your two sequences.

To ferret out this important information, you have to make an alignment. There are two kinds of alignments: the global alignments (where the two sequences are aligned along their entire lengths) and the local alignment (where the program only aligns the most similar portions of the two sequences).

Global alignments are usually easier to interpret than local ones. However, it only makes sense to build a global alignment if your sequences are related over their entire lengths — and if they don't contain very long insertions or deletions. If you don't know how to choose between global and local alignments, always go for the local methods.

TIP

When you're comparing two sequences, start making a dot plot of each sequence against itself. This approach makes it easier to identify potential repeats within each sequence. With this information in hand, interpreting the dot plot of the two sequences is much simpler.

A DIY guide to dot plots

Making a dot plot is much like having your sequences play tic-tac-toe against each other. For instance, imagine that you want to compare these two sequences:

```
Sequence 1: THEFATCAT
Sequence 2: THEFASTCAT
```

Telling them apart isn't so obvious — they are almost identical, except for the extra *S* in Sequence 2. (Of course, that difficulty gets much worse with real sequences.)

To make a dot plot, take a piece of paper, draw a grid, and write out Sequence 1 along the top with one letter above each column. Next, write out Sequence 2 on the left side with one letter next to each line, as in the accompanying matrix.

This is where the game begins. Consider each cell one by one, and cross it if the top sequence and the side sequence contain the same symbol. When this is done, you get something very much like what's shown here.

Of course, as in any tic-tac-toe game, the longest diagonals win! In this case, they indicate segments of similarity between your two sequences. The matrix below also makes it very easy to see that the two sequences are homologous over their entire length, with the exception of an extra residue in the vertical sequence.

```
..T H E F A T C A T
T X         X     X
H   X
E     X
F       X
A         X     X
S
T X          X    X
C              X
A         X      X
T X          X    X
```

The main difficulty when using the methods we list in Table 8-1 is the proper choice of parameters. Unfortunately, there is no absolute rule; producing the exact comparison you need is mostly a trial-and-error process. In the remainder of this chapter, we show you how to play with these parameters until they produce a result you can be satisfied with.

Making a Dot Plot

Dot plots are the simplest means of comparing two sequences. In fact, the dot plot — the kind you make using pencil and paper — is the only type of sequence comparison that you could do in the days when no one had true computer access. (See the sidebar "A DIY guide to dot plots.") Dot-plot techniques are closely related to sliding window methods — where a *window* is a portion of your sequence that you compare with other sequence portions of similar size. (For more on sliding windows, see Chapter 6). Dot plots have two major advantages: They're simple and don't require any biological hypothesis. Result: It's hard to make a mistake when making a dot plot. If you aren't sure what a dot plot really is, a quick glance at the matrix shown in the sidebar "A DIY guide to dot plots" may give you an idea.

The interpretation of a dot plot relies on the most sophisticated detection device available to you: your eyes. When it comes to identifying patterns such as diagonals, nothing beats the human eye, even when the signal is noisy. This is so true that any time you see a good signal on a dot plot, you can bet your life something interesting is going on — even if the statistics are too "smart" to see anything!

Choosing the right dot-plot flavor

In this section, we introduce a very powerful tool named Dotlet. Dotlet is very convenient because it's easy to use, free of charge, doesn't need any installation, and runs on (almost) any computer that has access to the Internet.

Dotlet is ideal for sequences with lengths of less than 10,000 amino acids or nucleotides. Thus it can be helpful for most proteins but is restricted to small DNA sequences. If you want to look at longer sequences online, you will have to use Dnadot (Table 8-2), a tool designed for fast dot-plotting between long sequences.

If your sequences are longer than 100,000 characters, no adequate tool is available online — you need to install a more powerful version of Dotlet on your computer. At least two such packages are available for free: Dotter and Dottup. You can find their respective specifications (and URLs) in Table 8-2.

Table 8-2 Various Flavors of Dot-Plot Programs

Name	*Used For*	*Range*	*URL*	*Platforms*
Dotlet	Proteins, DNA	10,000	`www.ch.embnet.org`	All (Java)
Dnadot	Proteins, DNA	100,000	`arbl.cvmbs.colostate.edu/molkit/dnadot/`	All (Java)
Dotter	Proteins, DNA	100,000	`www.cgr.ki.se/cgr/groups/sonnhammer/Dotter.html`	Unix, Linux, Windows
Dottup	Complete genomes, DNA	>100,000	`www.emboss.org`	Unix, Linux

Using Dotlet over the Internet

Dotlet is one of the most user-friendly dot-plot programs available over the Internet. It is a Java applet that you can access on the EMBnet server. An applet is a little program that loads automatically in your Internet browser and then runs on your own machine. The nice thing about it is that you don't need to install anything, as long as you have Java running (which most Internet browsers do by default these days).

Dotlet isn't a complicated program, but it contains many features. If you don't use these features properly, you may pass very close to the result you need but miss it. The general rule when using Dotlet is that you want to find a clear signal — and that you should keep playing with the parameters until you get one.

In the upcoming steps list, we show you some examples of what we mean by the term *clear signal* and give some tips on how best to obtain one.

Downloading Dotlet

You can download Dotlet with one click of a mouse, simply by pointing your browser to `www.isrec.isb-sib.ch/java/dotlet/Dotlet.html`. When you open this page, your browser also downloads the Dotlet program. This may take a few minutes if you're on a slow connection.

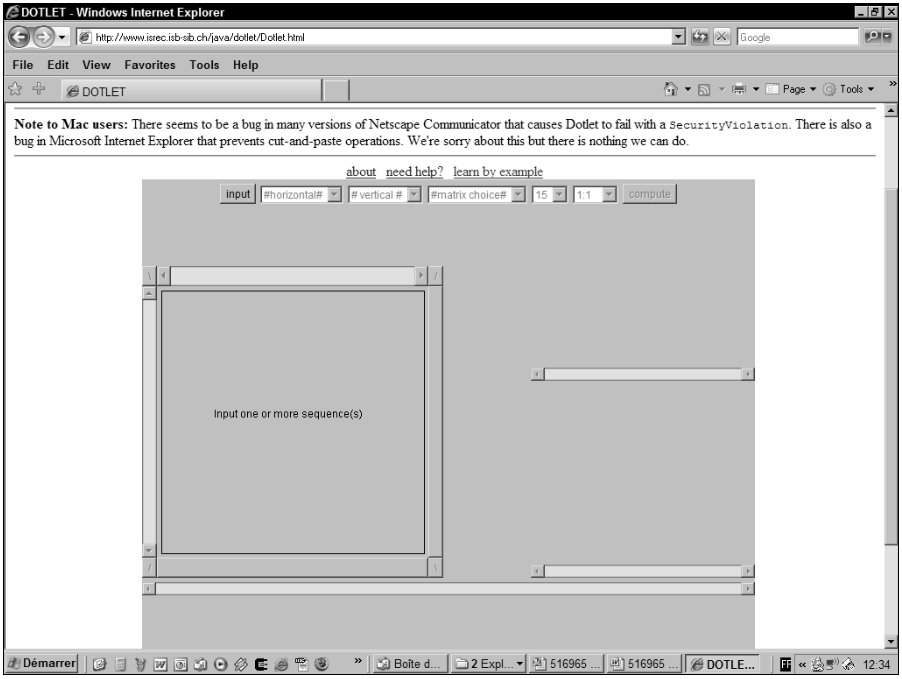
When the download is finished, your browser displays a page like the one shown in Figure 8-1.

If your browser does not display this page properly, it means that it can't run Java; you may need to update to the latest version of Internet Explorer or Netscape in order to get Java running on your computer.

Dotlet is not a server but a Java applet. Everything Dotlet does, it does on your own computer — not on the EMBnet server. For you, this has three consequences:

- You can use Dotlet offline after you download it once. This is useful if you work on a laptop that isn't always connected.
- The speed of the program depends on your own computer. The faster your computer, the faster Dotlet runs. Differences become apparent with sequences longer than 1,000 residues.
- When you use Dotlet, *you are not sending your sequences anywhere!* Each sequence you compare with Dotlet stays on your computer and isn't sent to the EMBnet server. If you work in a company that's at all interested in protecting proprietary information, you know this is a VERY GOOD thing!

  To make sure Dotlet isn't communicating with any server when you use it, choose File⇨Work Offline from your browser menu after you've downloaded Dotlet.



Note to Mac users: There seems to be a bug in many versions of Netscape Communicator that causes Dotlet to fail with a `SecurityViolation`. There is also a bug in Microsoft Internet Explorer that prevents cut-and-paste operations. We're sorry about this but there is nothing we can do.

Input one or more sequence(s)

Figure 8-1: The Dotlet page.

Entering your sequence in Dotlet

In the following steps list, we compare two sequences that contain a kinase domain. We're going to assume that you've already got Dotlet up and running in your browser (if not, the address is `www.isrec.isb-sib.ch/java/dotlet/Dotlet.html`):

1. **Open a new browser window by pressing Ctrl+N on your keyboard.**

   You need access to two browser windows, which is why we start with this step.

2. **Type the address** `www.expasy.ch/cgi-bin/get-sprot-fasta?P05049` **into the new window you opened in Step 1 and then press Enter or Return.**

   Type in the exact URL; do not omit the `?`. This URL gives you a direct access to Swiss-Prot if you know the ID number of your sequence. For our example, the ID number is P05049.

   After a bit of a wait, a new window appears, displaying the sequence associated with the ID number you just entered.

3. **Copy the sequence displayed in your browser using Ctrl-C.**

4. **Return to your first browser window (the Dotlet window shown in Figure 8-1) and click the Input button at the top left.**

   A pop-up window like the one in Figure 8-2 appears.

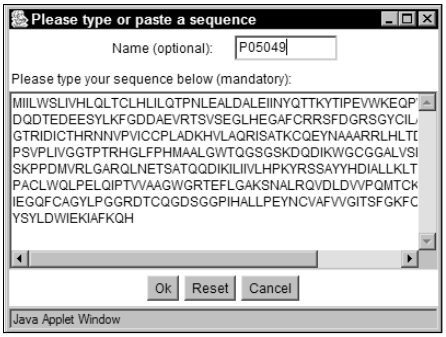
5. **Use Ctrl+V to paste your sequence into the pop-up window and then enter a title.**

   You must enter your sequence in raw format (only the sequence) as shown in Figure 8-2.

   The Name box is where you can enter the title of your sequence on the dot plot if you so desire.

6. **Press the OK button in the Sequence Window.**

   When you press the OK button, the window disappears; this means your sequence is now loaded in Dotlet.



Please type or paste a sequence

Name (optional): P05049

Please type your sequence below (mandatory):

```
MIILWSLIVHLQLTCLHLILQTPNLEALDALEIINYQTTKYTIPEVWKEQP
DQDTEDEESYLKFGDDAEVRTSVSEGLHEGAFCRRSFDGRSGYCIL
GTRIDICTHRNNVPVICCPLADKHVLAQRISATKCQEYNAAARRLHLTD
PSVPLIVGGTPTRHGLFPHMAALGWTQGSGSKDQDIKWGCGGALVSI
SKPPDMVRLGARQLNETSATQQDIKILIIVLHPKYRSSAYYHDIALLKLT
PACLWQLPELQIPTVVAAGWGRTEFLGAKSNALRQVDLDVVPQMTCK
IEGQFCAGYLPGGRDTCQGDSGGPIHALLPEYNCVAFVVGITSFGKFC
YSYLDWIEKIAFKQH
```

Figure 8-2: The sequence window in Dotlet.

If you want to compare a sequence only with itself, go directly to Step 13.

7. **Type the address** `www.expasy.ch/cgi-bin/get-sprot-fasta?P08246` **into the new window you opened in Step 1, and press Enter or Return.**

   This is the second sequence you want to load into Dotlet.

8. **Copy the sequence displayed in your browser.**

9. **In the Dotlet window (refer to Figure 8-2), click the Input button.**

10. **Paste your sequence using Ctrl-V in the pop-up window that appears and enter a title.**

11. **Click the OK button in the pop-up Window.**

    When you click the OK button, the window disappears; this only means that your sequence is now loaded in Dotlet — it hasn't disappeared for good!

12. **Back in your original browser window — the Dotlet window — choose the two sequences you want to compare from the two drop-down windows nearest to the Input button, as shown in Figure 8-3.**

    Figure 8-3 shows you how to use the pull-down menus to choose the sequences you want to compare. The names in the menus are those you provided in the Title box in Steps 5 and 10.

    - The sequence you choose in the first pull-down menu is the horizontal sequence (top of the dot plot, left to right).
    - The sequence you choose in the second menu is the vertical sequence (left side of the dot plot, top to bottom).

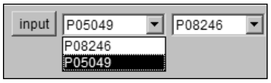


Figure 8-3: The sequence pull-down menus in Dotlet.

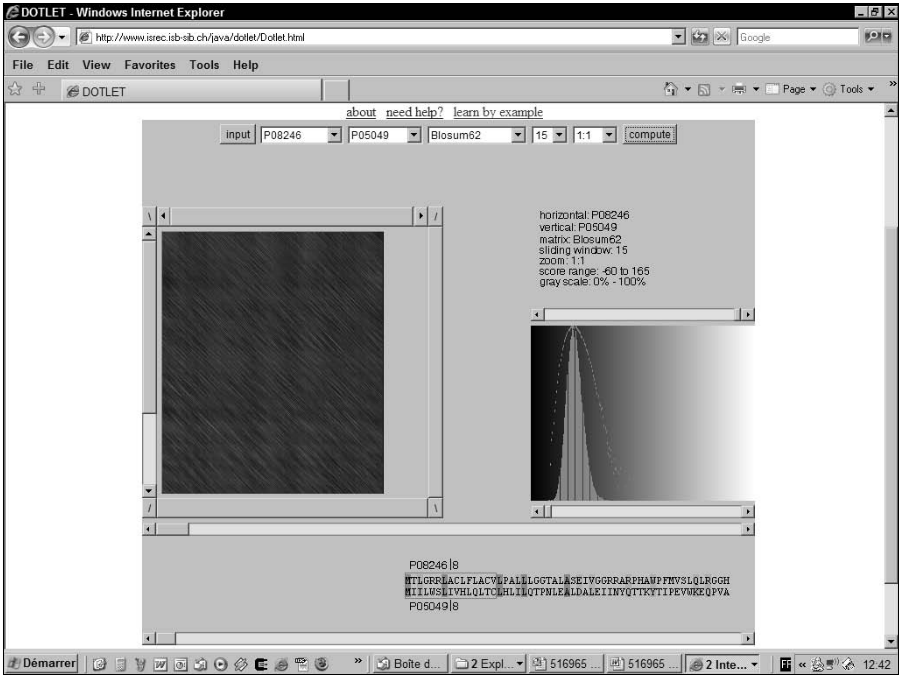
13. **In the Dotlet window, click the Compute button at the far right.**

    Dotlet never takes action automatically (except when you change the threshold). Any time you change one of its parameters, you must inform Dotlet by clicking the Compute button.

    The output looks like what's shown in Figure 8-4. This figure shows Dotlet's three windows:

    - **The dot-plot window:** The large window on the left contains your dot plot.
    - **The threshold window:** This small window on the right is the one you must use to tune the sensitivity of your dot plot.
    - **The alignment window:** This is the long narrow window below that shows which pair of residues from the two sequences corresponds to each dot in the dot-plot window.

The default output of Dotlet is usually difficult to interpret. To make it talk some sense, you have to refine it — which we show you how to do in the next section.


DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
File Edit View Favorites Tools Help
DOTLET
about need help? learn by example
input
compute
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
12:42

P08246 | P05049 | Blosum62 | 15 | 1:1

```
horizontal: P08246
vertical: P05049
matrix: Blosum62
sliding window: 15
zoom: 1:1
score range: -60 to 165
gray scale: 0% - 100%
```

```
P08246|8
MTLGRRLACLFLACVLPALLLGGTALASEIVGGRRARPHAWPFMVSLQLRGGH
MIILWSLIVHLQLTCLHLILQTPNLEALDALEIINYQTTKYTIPEVWKEQPVA
P05049|8
```

Figure 8-4: Default output of Dotlet.

Fine-tuning your Dotlet

The sequence of instructions we show here isn't necessarily optimal, but it follows a top-down approach. We start with settings that you can roughly adjust, and finish with settings that you must fine-tune to yield an informative dot plot.

1. **Set the zoom factor.**

   By default, Dotlet has a zoom factor of 1:1 — which means that 1 pixel corresponds to 1 residue.

   If your sequences are long, you may not see the entire dot plot. This is exactly what happens in Figure 8-4. You can tell this is so because the scroll bar down the left side is *active* — you'd need to use the rectangular box in the scroll bar to scroll down to see the rest of the display. To see the entire dot plot, do the following:

   a. **In the Zoom selector — the pull-down menu to the left of the Compute button — choose 1:2, as shown in Figure 8-5.**

   b. **Click the Compute button.**

      Dotlet now displays a complete dot plot of your two sequences.

   Seeing the overall picture is useful, even if it looks a bit small. After you get a feeling for the regions that are interesting in your dot plot, you can always zoom back.

2. **Choose the window size.**

   When Dotlet compares two amino acids — or nucleotides — it also compares the amino acids — or nucleotides — in the immediate vicinity. The window size controls the size of this neighborhood (for some helpful hints, see the “Getting the right window size” sidebar). To use our example, you have to change the default value as follows:

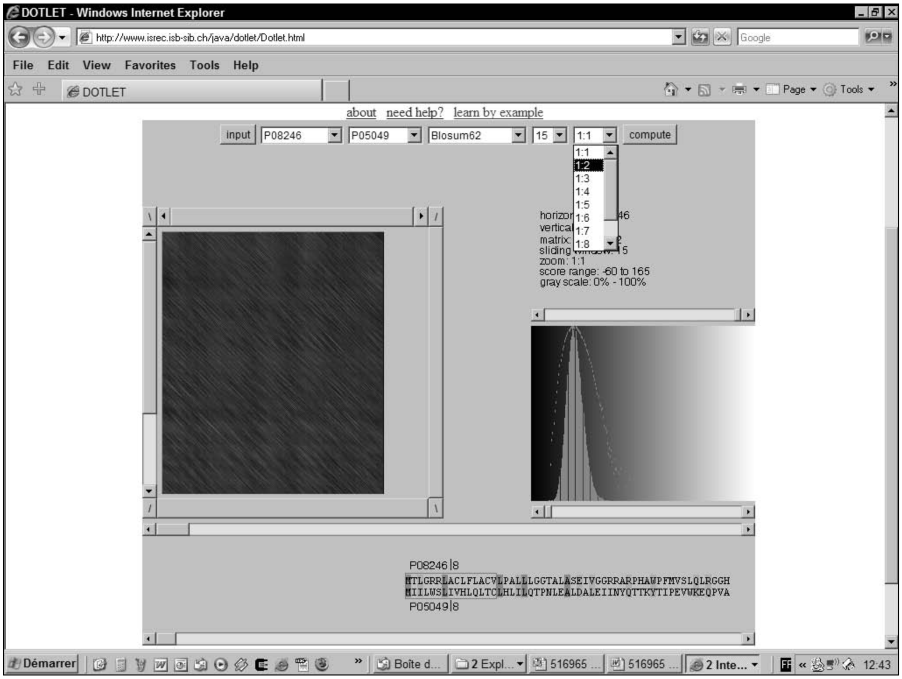
   a. **Choose 51 from the Window pull-down menu (the menu to the left of the Zoom selector).**

   b. **Click the Compute button.**

   After changing the window size, you MUST also tune the threshold (see Step 3). There is little point in changing the window size if you don’t adjust the threshold accordingly.

3. **Adjust the threshold.**

   The *threshold* is a value that defines the color of each dot in the dot-plot window. Dots with a score above the threshold are white; those with a score below the threshold are black. The threshold is by far the most powerful and the most delicate parameter when you’re making a dot plot. Fortunately, Dotlet lets you adjust the threshold dynamically and see the effect in real time. (You don’t even have to click the Compute button!)


DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
Google
File Edit View Favorites Tools Help
DOTLET
Page Tools
about need help? learn by example
input
P08246
P05049
Blosum62
15
1:1
compute
1:1
1:2
1:3
1:4
1:5
1:6
1:7
1:8
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
FR
12:43

horizon 46
vertical
matrix: 2
sliding 15
zoom: 1:1
score range: -60 to 165
gray scale: 0% - 100%

```
P08246|8
MTLGRRLACLFLACVLPALLLGGTALASEIVGGRRARPHAWPFMVSLQLRGGH
MIILWSLIVHLQLTCLHLILQTPNLEALDALEIINYQTTKYTIPEVWKEQPVA
P05049|8
```

Figure 8-5: Choosing the right zoom factor in Dotlet.

Getting the right window size

Here are some handy rules for making your window the right size:

- Long windows make clean dot plots. They make sure that two similar amino acids (or nucleotides) only yield a dot if the amino acids (or nucleotides) around them are also similar. Gurus say that long windows make the dot plot more *stringent.* Of course, being too stringent is just as bad as not being stringent enough. In order to achieve the right equilibrium, you can use the following simple guidelines:
- The size of a window should be within the same range as the size of the elements you're looking for. For instance, if you're looking for conserved domains in proteins, a size of 50 amino acids or higher is appropriate.
- Shorter windows are more sensitive but bring some noise with them. They may help if you're looking at protein domains that are very distantly related.
- It's convenient to start with a large window and narrow it a little until the signal you're looking for appears.

On the top of the threshold window — the window on the right side for the Dotlet display in Figure 8-4 — there is a little horizontal scroll bar with a cursor. This is the *top cursor.* On the bottom of this same window, there is another small horizontal scroll bar with another small cursor. This is the *bottom cursor.*

The horizontal axis in the threshold window represents the score. Low scores are to the left, high scores to the right. The big peak on the curve shows you the most common score for two residues chosen randomly.

You can imagine the threshold as two vertical lines going through the top and the bottom cursor. Its position controls the appearance of the dot-plot window (the large window on the left in Figure 8-4). In this dot-plot window

- Dots with a score below the bottom cursor threshold are black.
- Dots with a score above the top cursor threshold are white.
- Dots with a score between the two cursors' thresholds are gray.

The game is all about setting the threshold so interesting residues appear as white dots on a black background in the dot-plot window on the left. You can use the following procedure to achieve this effect:

a. **Drag the bottom cursor in the threshold window entirely to the right.**

   This makes the dot plot on the left entirely black.

b. **Grab the right side of the Threshold Window by the middle, and slowly drag it to the left, while holding your right mouse button pressed down.**

While you drag, keep an eye on the dot plot to see a conserved segment appear, as shown in Figure 8-6.

TIP

When the top cursor and the bottom cursor are on the top of each other, dots in the dot-plot window are either black or white. If you shift one of the cursors to the left or to the right, the dots whose scores are between these two thresholds appear in gray. This can help smooth the plot and make it look nicer for publication.

TIP

If you want to see black dots on a white background, shift the top cursor slightly to the right of the bottom cursor. This inverts the color chart.

4. **Save your dot plot.**

Saving your results is the most delicate part of using Dotlet. Not all browsers can print these results — and we didn't find one that could save the display as a file. As far as we know, there are only two ways to keep a record of your display:

Print the browser content into a PDF file. To do so, you must have a PDF printer installed on your computer.

Use the ol' screen-capture technique. You can use it with any graphic display that gives you trouble when trying to save. Follow this procedure to do a screen capture.

a. **Press the PrntScrn button on your keyboard (usually top right).**

b. **Start a Power Point presentation.**

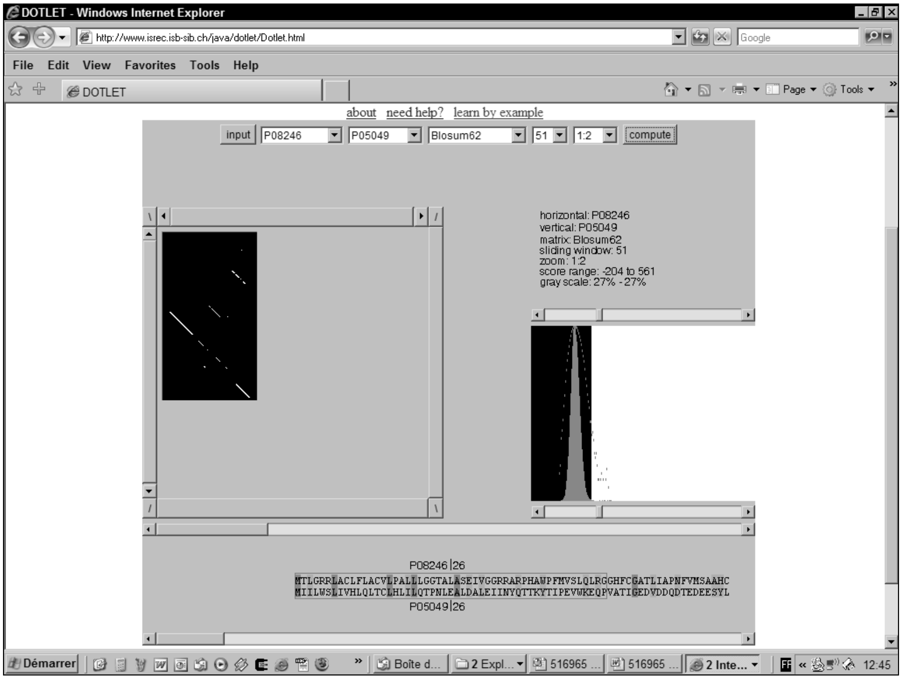
c. **Press Ctrl+V to cut and paste your screen-capture into the presentation.**

You can now print your dot plot or include it in any type of document.

Interpreting your results

P05049 and P08246, the two proteins we chose for the previous steps list (in the section "Entering your sequence in Dotlet"), are very distantly related. If you use BLAST to find one with the other, you get a match with an E-value of 10-^{4}. If you have been through Chapter 7 — which introduces you to the intricacies of BLAST — you already know that getting really excited over such a match isn't a good idea. It could even be a false positive.

Knowing that P05049 is a serine protease, it would be interesting to check the protease activity of P08246 in the wet lab. The problem is that when you launch a costly experiment on the basis of such a weak BLAST match, you're definitely taking a chance. If you fail, the probabilities are high that your boss will ask the lethal question, "What in the world where you thinking?!"



DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
File Edit View Favorites Tools Help
DOTLET
about need help? learn by example
input
compute
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
12:45

P08246 | P05049 | Blosum62 | 51 | 1:2

horizontal: P08246
vertical: P05049
matrix: Blosum62
sliding window: 51
zoom: 1:2
score range: -204 to 561
gray scale: 27% - 27%

```
P08246|26
MTLGRRLACLFLACVLPALLLGGTALASEIVGGRRARPHAWPFMVSLQLRGGHFCGATLIAPNFVMSAAHC
MIILWSLIVHLQLTCLHLILQTPNLEALDALEIINYQTTKYTIPEVWKEQPVATIGEDVDDQDTEDEESYL
P05049|26
```

Figure 8-6: Dotlet with a proper parameter setting.

This is the reason why — before doing this experiment — it would be reassuring to make sure that P05049 and P08246 are genuinely homologous in the serine protease region.

Making a dot plot is a good way to answer this question. A dot plot like the one we produced in the previous steps list is very much to the point. Figure 8-6 makes a good case for the existence of a conserved domain between the two proteins. You can make sure of this by looking at their Swiss-Prot annotation (accessible at `www.expasy.ch`). Swiss-Prot says that P05049 and P08246 both contain a serine protease domain in the region that the dot plot indicates.

Of course, in real life you may not have a Swiss-Prot annotation for the two proteins. In fact, in the real world, you may not have any annotation at all! That should not deter you from making a dot plot. Even if you don't know its function, the identification of a conserved domain between two protein sequences is a good way to locate the specific region responsible for a function.

Doing biological analysis with a dot plot

The preceding example shows you how to spot a distantly related domain shared by two proteins. Dot plots give very characteristic representations of biological relationships. Using them is mostly a matter of recognizing a few graphic patterns and knowing which situations they correspond to.

In the following example, we show you some of these situations. For each one, we assume that you have entered your sequences in Dotlet by using a procedure like the one we describe in the section “Using Dotlet over the Internet.”

Identifying tandem repeats

Proteins commonly contain a small domain repeated many times over. It seems that internal duplication is a tool often used by evolution to create new proteins or make them function more efficiently. Domains that you may often find duplicated include the Fibronectin domain (we like to refer to it as “molecular Velcro”), EF-hands that bind calcium, or Zn fingers involved in DNA binding.

Dot plots constitute by far the best way to spot repeated domains. In Figure 8-7, we show you the YE73 human protein sequence. YE73 is a potential human transcription factor (involved in RNA transcription) that contains 13 very conserved Zn finger domains in tandem.

To analyze this sequence, follow these steps:

1. **Point your browser to** `www.expasy.ch/cgi-bin/get-sprot-fasta?Q9P255`**.**

   Q9P255 is the ID number for the YE73 human protein sequence.

2. **Enter the sequence into Dotlet, following the procedure we outline in the section “Using Dotlet over the Internet.”**

3. **Set the threshold and window settings so they match those in Figure 8-7.**

   You want a window size of 21, a zoom of 1:2, and a gray scale for the top and bottom cursors of the threshold window set to 39%.

   The pattern in Figure 8-7 is typical of tandem repeats. Notice the following features:

   - The main diagonal represents the sequence against itself.
   - Repeats appear as long continuous diagonals above and below the main diagonal.
   - The diagonals are evenly spaced.
   - The collection of diagonals is bound by the shape of a square.

Anytime you see this type of graphic pattern, you know you’re looking at tandem repeats. If this is the case, you can use this collection of diagonals to make three interesting deductions:

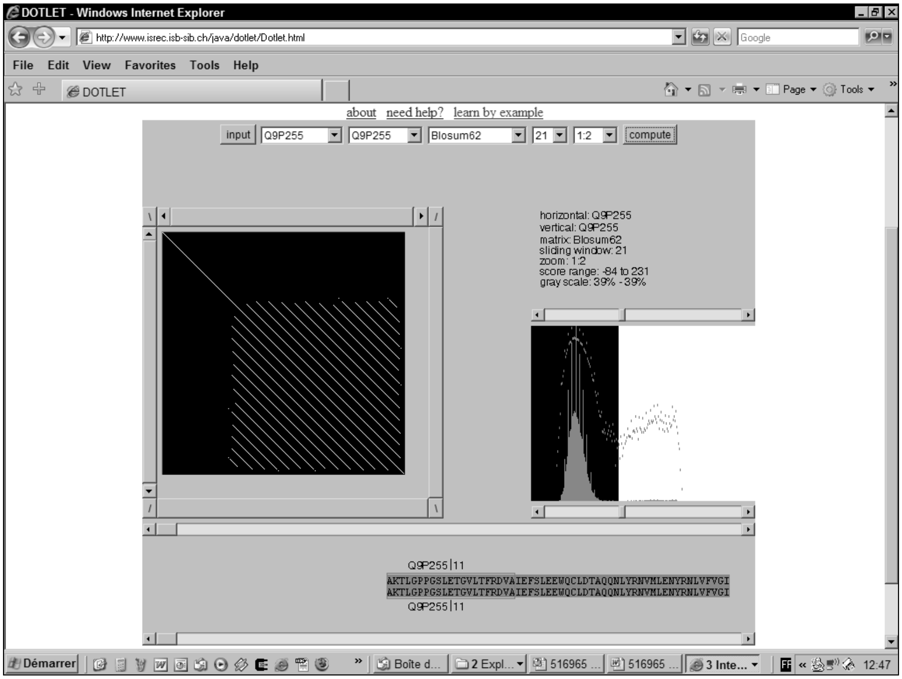
- The number of repeats is equal to the number of diagonals above (or below) the main diagonal (including the main diagonal).
- The distance between two adjacent diagonals represents the size of the repeat if no spacing falls between the repeats.
- The shortest diagonal gives you the coordinates of a single repeat unit.

YE73 is an extreme example because it contains highly conserved repeats. Life isn't always so simple — and, in most cases, the repeated domains you come across aren't so similar (because of accumulated mutations). Proteins that contain distantly related repeats produce more complicated dot-plot patterns. In these proteins, each repeat unit doesn't necessarily recognize every other unit.

Figure 8-8 shows you such an example. Tf3a is a transcription factor that also contains Zn finger domains in tandem. You can find the sequence at

```
www.expasy.ch/cgi-bin/get-sprot-fasta?P03001
```

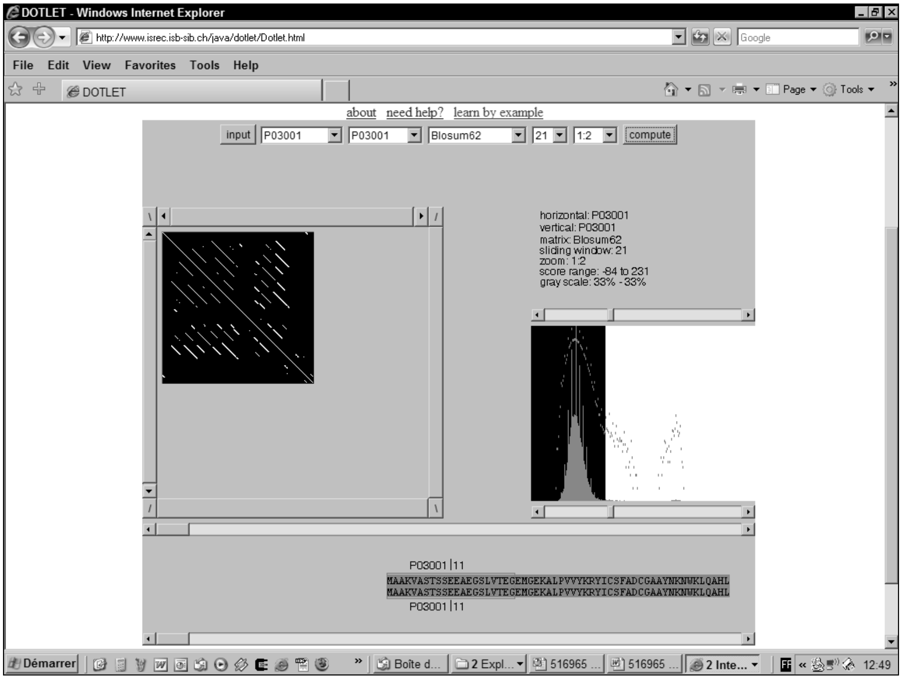
Tf3a tells the story of Zn finger domains that have been duplicated and have diverged since then. These Zn fingers are less conserved, and show a pattern much less regular than what we obtained with YE73.


DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
Google
File Edit View Favorites Tools Help
DOTLET
Page
Tools
about need help? learn by example
input
Q9P255
Q9P255
Blosum62
21
1:2
compute
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
3 Inte...
12:47

horizontal: Q9P255
vertical: Q9P255
matrix: Blosum62
sliding window: 21
zoom: 1:2
score range: -84 to 231
gray scale: 39% - 39%

```
Q9P255|11
AKTLGPPGSLETGVLTFRDVAIEFSLEEWQCLDTAQQNLYRNVMLENYRNLVFVGI
AKTLGPPGSLETGVLTFRDVAIEFSLEEWQCLDTAQQNLYRNVMLENYRNLVFVGI
Q9P255|11
```

Figure 8-7: Looking at tandem repeats with Dotlet.


DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
Google
File Edit View Favorites Tools Help
DOTLET
Page
Tools
about need help? learn by example
input
compute
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
12:49

P03001 | P03001 | Blosum62 | 21 | 1:2

horizontal: P03001
vertical: P03001
matrix: Blosum62
sliding window: 21
zoom: 1:2
score range: -84 to 231
gray scale: 33% - 33%

```
P03001 |11
MAAKVASTSSEEAEGSLVTEGEMGEKALPVVYKRYICSFADCGAAYNKNWKLQAHL
MAAKVASTSSEEAEGSLVTEGEMGEKALPVVYKRYICSFADCGAAYNKNWKLQAHL
P03001 |11
```

Figure 8-8: Looking at the poorly conserved tandem repeats of Tf3a with Dotlet.

Tandem domains are not the only form of repeated domains in proteins or DNA. It is safe to assume that any kind of repetition you can think of probably occurs in proteins or in nucleic acids. If it has not yet been found, you can bet it's about to be discovered!

The signature of a repetition is always the same: It's a diagonal that lies off the main diagonal when you compare the sequence with itself.

Repeated domains can help you elucidate the function of your protein. If you find a repeated domain with no known function — and no similarity with any characterized protein — you can do the following:

1. **Extract each repeat unit.**
2. **Make a multiple alignment of the domains.**

   See Chapter 9 for more about how to do this.
3. **Identify conserved positions in the domain.**
4. **Turn your domain into a PROSITE pattern or a profile.**

   See Chapter 6 for more on PROSITE patterns.
5. **Scan Swiss-Prot in order to check whether this pattern is associated with a function.**

Finding low-complexity regions in proteins

Low-complexity regions are segments that contain a certain amino acid more often than you would expect in *normal* proteins. For instance, a segment of 100 amino acids that contains 45 prolines is definitely a low-complexity segment.

Low-complexity regions often have biological functions like protein-protein interaction (Leucine Zipper) or nonspecific DNA/RNA binding (ARG-rich domains).

The good news about low-complexity regions is that seeing them on a dot plot is really easy. When you compare a sequence with itself, low-complexity regions pop up as squares. You can see one of these in Figure 8-9.

To reproduce the results you see on Figure 8-9, follow these steps:

1. **Point your browser to** `www.expasy.ch/cgi-bin/get-sprot-fasta?P21997`**.**

   P21997 is the ID number for the sulfated surface glycoprotein 185.

2. **Enter the sequence into Dotlet, following the procedure we outline in the section "Using Dotlet over the Internet."**

3. **Set the settings so they match those in Figure 8-9.**

   You need a window setting of 7, a zoom setting of 1:2, and a gray scale for the top and bottom cursors of the threshold window set to 44%.

Analyzing nucleic acids with Dotlet

By nature, dot plots are well suited for mapping genes. Unfortunately, Dotlet can be a bit slow for such applications, which routinely require the comparison of sequences longer than 10,000 residues.

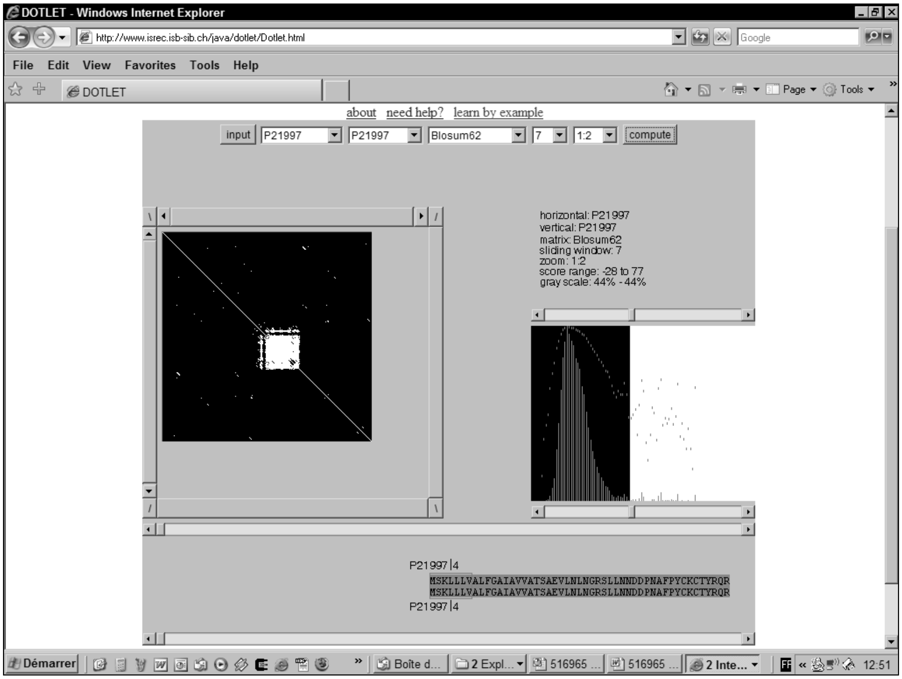
TIP

Dotlet contains two useful features for analyzing nucleotide sequences:

- **If you compare a protein sequence and a nucleotide sequence:** Dotlet automatically translates the nucleotide sequences into its three possible protein-reading frames. This way you can pinpoint exon/intron boundaries.
- **If you compare a nucleotide sequence with itself:** Dotlet automatically replaces one of the sequences with its complementary sequence. This way, if your sequence contains two complementary strands (like a hairpin stem in an RNA structure), they appear as diagonals going from the top-right corner toward the bottom-left (perpendicular to the main diagonal).

If you want to use Dotlet for more refined sequence analysis, we recommend the excellent online tutorial produced by Dotlet authors, available at

```
www.isrec.isb-sib.ch/java/dotlet/dotlet_examples.html
```



DOTLET - Windows Internet Explorer
http://www.isrec.isb-sib.ch/java/dotlet/Dotlet.html
File Edit View Favorites Tools Help
DOTLET
about need help? learn by example
input
compute

P21997 P21997 Blosum62 7 1:2

```
horizontal: P21997
vertical: P21997
matrix: Blosum62
sliding window: 7
zoom: 1:2
score range: -28 to 77
gray scale: 44% - 44%
```

```
P21997 |4
    MSKLLLVALFGAIAVVATSAEVLNLNGRSLLNNDDPNAFPYCKCTYRQR
    MSKLLLVALFGAIAVVATSAEVLNLNGRSLLNNDDPNAFPYCKCTYRQR
P21997 |4
```

Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
12:51

Figure 8-9: Identifying low-complexity segments with Dotlet.

Making Local Alignments over the Internet

If you've gone through the dot-plot section of this chapter, you know that making a dot plot is an ideal way to get an overview of the relationship between your sequences. Nonetheless, dot plots aren't predictive methods: They show a signal but don't tell you what it means. To make this interpretation, you need a method that produces an alignment.

REMEMBER

There are two kinds of alignments: *global* (where the two sequences are aligned over their entire lengths) and *local* (where the program only aligns the most similar portions of your two sequences and ignores the rest).

TIP

If you have a dot plot showing that your two sequences are related over their entire lengths, you may not need to make a local alignment, and you can go directly to the global alignment section of this chapter. If you're not sure, stay here!

There are two major reasons for analyzing your sequences with local alignments:

- To compare two distantly related sequences that share only a few noncontiguous domains
- To analyze repeated elements within a single sequence

Local alignment methods do exactly what their name implies: You give them two sequences, and they output an alignment of the two most similar portions of these sequences.

The nice thing about local alignment methods is that they automatically get rid of the amino acids or nucleotides that they can't align.

In theory, a local alignment should correspond to one of the diagonals that appear on the dot plot. In practice, however, things aren't so simple; you may see a signal on the dot plot that doesn't correspond to any local alignment, and sometimes local alignment programs report alignments that don't appear on the dot plot. Only when the dot plot and the alignment agree perfectly can you be sure.

Another rule with sequence alignments is that you should trust them only when they are *clearly correct.* (In this section, we show you what we mean by "clearly correct.") The only situation where you can use a suspicious alignment is when structural or experimental information exists that supports this alignment.

Choosing the right local-alignment flavor

Two types of methods exist for making local pairwise alignments: a fast, heuristic method named BLAST and a slower, more accurate one named Lalign. Table 8-3 lists the pros and cons of these two methods so you can select the one that suits you best.

The version of BLAST that compares two sequences is the same one you can use for searching a database. It's adapted so you can restrict it to two sequences only, but it doesn't generate alignments different from those BLAST reports when it's searching complete databases. In short, if you want to check a hit, BLASTing two sequences can't tell you anything more than the database search. You can access this special flavor of BLAST at

```
www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi
```

Table 8-3 Using BLAST or Lalign for a Pairwise Comparison

Property	*BLAST*	*Lalign*
Speed	Very fast	Slower
Size of sequences you can use	Very long sequences	Shorter sequences
Score	E-value	Matrix score and E-value
Alignments	Reports the single best	Reports the ten best (or more)
Sequence type	Best with DNA	Best with proteins

Using Lalign to find the ten best local alignments

The main problem when using BLAST is that this program only returns one alignment: the best-scoring one. This is very useful for searching databases, but it may not be so interesting if the region of the sequences you're interested in happens to belong to a *good* local alignment but not to the *best* local alignment. It's a bit like a music store that would only sell you one CD — the number-one bestseller of the week.

Lalign is the ideal complement to BLAST. It is slower but more accurate, and it returns as many local alignments as you want (the best scoring one, the second best scoring, and so on, up to a number that you specify). Don't worry about the speed issue. On sequences shorter than a thousand symbols, you may not notice any difference between the bl2seq server and Lalign.

In general, Lalign is great when it comes to analyzing complicated proteins full of repeats. In the following steps list, we use Lalign to extract local alignments from two distantly related sequences that both contain a serine protease domain:

1. **Point your browser to** `www.ch.embnet.org/software/LALIGN_form.html`**.**

   The Lalign page of `ch.EMBnet.org` appears.

2. **Choose Local from the alignment method options, as shown in Figure 8-10.**

   You can use the Lalign interface to do local and global alignments.

3. **Choose the number of reported subalignments.**

   You should choose a number that corresponds roughly to the number of diagonals you observed on the dot plot. If you want Lalign to report 10 local alignments, select 10.

4. **Select the default substitution matrix.**

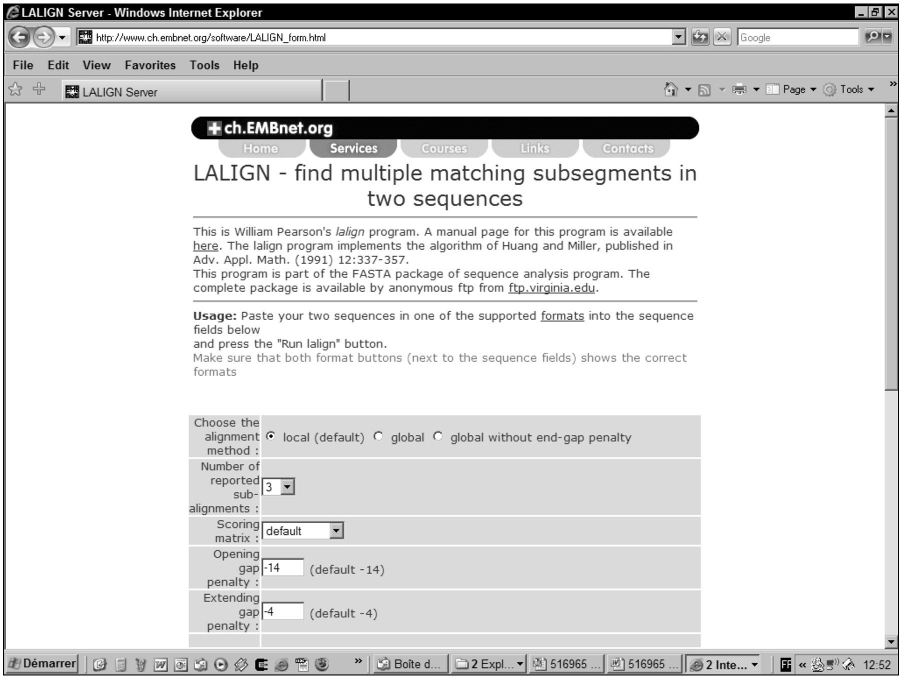
   The *substitution matrix* controls the cost of the mutations when Lalign aligns the sequences.

   The default matrix is an identity matrix for DNA and a BLOSUM45 for proteins. In most cases, this default is entirely appropriate. If you want to change the matrix, bear in mind that high BLOSUM indexes make Lalign more stringent — and result in shorter alignments. Low PAM indexes have the opposite effect.

5. **Set the gap-opening penalty.**

   The *gap-opening penalty* defines the cost for opening a gap in one of the sequences.

   If you set the gap-opening penalty higher than its default value, local alignments that contain gaps may be split into several shorter alignments. There is no simple rule to predict the optimal value for the gap penalty.



LALIGN - find multiple matching subsegments in two sequences

This is William Pearson's *lalign* program. A manual page for this program is available here. The lalign program implements the algorithm of Huang and Miller, published in Adv. Appl. Math. (1991) 12:337-357.
This program is part of the FASTA package of sequence analysis program. The complete package is available by anonymous ftp from ftp.virginia.edu.

Usage: Paste your two sequences in one of the supported formats into the sequence fields below
and press the "Run lalign" button.
Make sure that both format buttons (next to the sequence fields) shows the correct formats

Choose the alignment method :	● local (default) ○ global ○ global without end-gap penalty
Number of reported sub-alignments :	3
Scoring matrix :	default
Opening gap penalty :	-14 (default -14)
Extending gap penalty :	-4 (default -4)

Figure 8-10: The Lalign home page.

The default values have been tuned on the default matrix. If you change the matrix, Lalign does not automatically readjust the values. For instance, if you use a BLOSUM with an index lower than 45, you must decrease the gap-opening value and set it lower than 14.

6. **Set the gap-extension penalty.**

   The *gap-extension penalty* is an extra penalty proportional to the length of the gap. This penalty is added to the gap-opening penalty to yield the complete gap cost. It must be about ten times lower than the gap-opening penalty.

   When you compare distantly related sequences, a high gap-opening penalty and a very low gap-extension penalty often yield the best results. They indicate that gaps should be penalized more on the basis of their existence than of their length.

7. **Choose Swiss-Prot ID or AC from the Input Sequence Format pull-down menu.**

   You'll need to scroll down the Lalign page to get to the menu.

8. **Enter** P05049 **in the first sequence box.**

   P05049 is the accession number of the snake serine protease. If you want to use a sequence that has no accession number, paste it in the window — and make sure that you chose the correct format in Step 7.

9. **Set the second format selector to Swiss-Prot ID or AC.**

10. **Enter** P08246 **in the second sequence box.**

    P08246 is the accession number of the human leucocyte elastase.

11. **Click the Run Lalign button.**

    The computation should be relatively fast (less than one minute). If the program doesn't return any results, check to make sure you've submitted the right parameters.

    Lalign returns the results in the form of an HTML document.

12. **Save your results.**

    Use the File➪Save As option of your Web browser to keep this result in a file.

Interpreting the Lalign output

Lalign reports the number of local alignments you have specified — as in the preceding Step 3 — sorted according to their score. An interesting property of Lalign is that it only reports *nonoverlapping* alignments.

This means that, in the Lalign output, you can find two amino acids or nucleotides aligned together only once. They can appear in several alignments of course, but they appear in front of each other only once. This prevents Lalign from reporting trivial variations of a very good local alignment.

Figure 8-11 shows you the output of Lalign given the two sequences used in the preceding steps list. Lalign reports the alignments in a BLAST-like format, sorted according to their E-values. On the first line associated with each alignment are the following features:

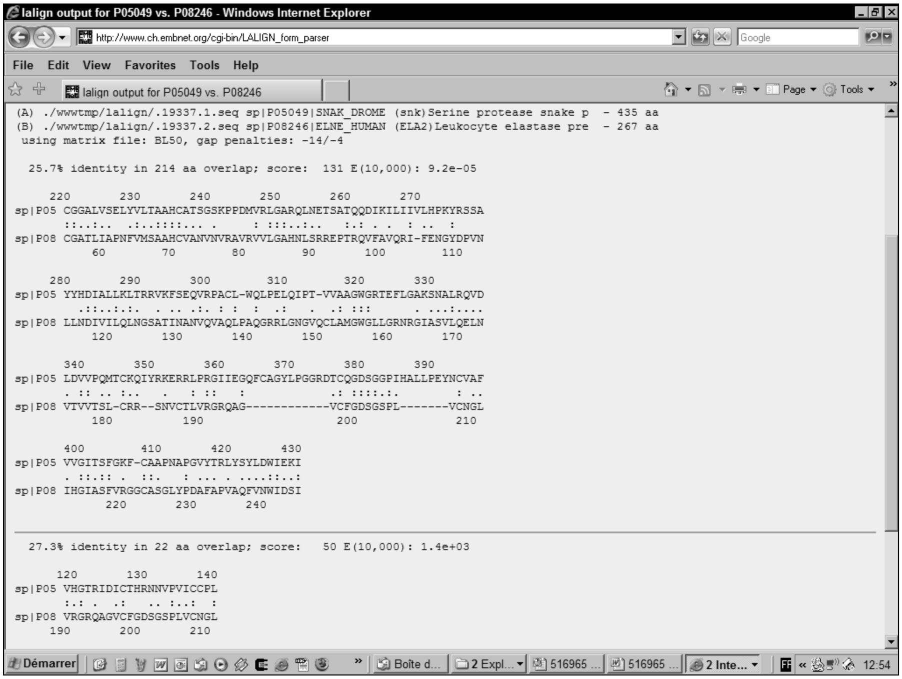
- **The percent identity:** The proportion of identical residues aligned with one another. For instance, you can see in Figure 8-11 that the best local alignment has a percent-identity score of 25.7 percent; the second-best score is 27.3 percent.
- **The local alignment length (Overlap):** This is the total length of the local alignment.
- **The score:** This score sums up the cost of the gaps and substitutions, as given by the substitution matrix and the gap penalties. Generally speaking, the higher the score, the better the alignment — yet be aware that the absolute value has no clear meaning. The E-value is a better indicator of the alignment's quality.
- **The E-value:** This value tells you how many times you could have expected to find such a good alignment by chance, given your two sequences. Be aware that this E-value is much less meaningful than the one BLAST reports when searching a database. A good E-value must be below 10-^{4}.

The alignment itself contains three types of information:

- **The residue index on the line above the sequence.** The residue corresponding to an index is the one below the last digit of this index.
- **The alignment itself, with gaps represented by dashes.**
- **Identity and similarity.** In the line between the two aligned sequences, the (_) symbol means identity, while the (.) symbols mean similarity. Two residues are similar when their substitution score is greater than 0.

The first alignment reported corresponds to the conserved serine protease domain. On its own, this alignment isn't really convincing: It contains less than 26 percent identity over about 200 residues.

To be really convinced, we would need a much higher similarity (at least close to 30 percent) and a better (lower) E-value. The reason we may trust this alignment is because it is consistent with the signal we previously saw on the dot plot. (Refer to Figure 8-7.) The fact that these two different analyses give compatible results (Dotlet and Lalign) makes a good case for the existence of a conserved serine protease domain in our proteins.


lalign output for P05049 vs. P08246 - Windows Internet Explorer
http://www.ch.embnet.org/cgi-bin/LALIGN_form_parser
Google
File Edit View Favorites Tools Help
lalign output for P05049 vs. P08246
Page
Tools
Démarrer
Boîte d...
2 Expl...
516965 ...
516965 ...
2 Inte...
FR
12:54

```
(A) ./wwwtmp/lalign/.19337.1.seq sp|P05049|SNAK_DROME (snk)Serine protease snake p  - 435 aa
(B) ./wwwtmp/lalign/.19337.2.seq sp|P08246|ELNE_HUMAN (ELA2)Leukocyte elastase pre  - 267 aa
 using matrix file: BL50, gap penalties: -14/-4

  25.7% identity in 214 aa overlap; score:  131 E(10,000): 9.2e-05

     220       230       240       250       260       270
sp|P05 CGGALVSELYVLTAAHCATSGSKPPDMVRLGARQLNETSATQQDIKILIIVLHPKYRSSA
       ::..:..  .:..::::... .     : :::..:..   :.: . .  : ..  :
sp|P08 CGATLIAPNFVMSAAHCVANVNVRAVRVVLGAHNLSRREPTRQVFAVQRI-FENGYDPVN
           60        70        80        90       100        110

     280       290       300        310        320       330
sp|P05 YYHDIALLKLTRRVKFSEQVRPACL-WQLPELQIPT-VVAAGWGRTEFLGAKSNALRQVD
         .::..:.:.  . .. .:. : :  :  .:   .  .: :::      . ...:....
sp|P08 LLNDIVILQLNGSATINANVQVAQLPAQGRRLGNGVQCLAMGWGLLGRNRGIASVLQELN
           120       130       140       150       160       170

       340       350       360       370       380       390
sp|P05 LDVVPQMTCKQIYRKERRLPRGIIEGQFCAGYLPGGRDTCQGDSGGPIHALLPEYNCVAF
       . :: .. :..   .   : ::   :            .: ::::.:.        : ..
sp|P08 VTVVTSL-CRR--SNVCTLVRGRQAG------------VCFGDSGSPL-------VCNGL
           180          190                 200              210

       400        410       420       430
sp|P05 VVGITSFGKF-CAAPNAPGVYTRLYSYLDWIEKI
       . ::.:: .  ::.   : ... . ....::..:
sp|P08 IHGIASFVRGGCASGLYPDAFAPVAQFVNWIDSI
             220       230       240

  27.3% identity in 22 aa overlap; score:   50 E(10,000): 1.4e+03

      120       130       140
sp|P05 VHGTRIDICTHRNNVPVICCPL
       :.: .  .:   .. :..:  :
sp|P08 VRGRQAGVCFGDSGSPLVCNGL
     190       200       210
```

Figure 8-11: Output of Lalign.

If you play with the Lalign parameters, you may find that stronger gap-opening or gap-extension penalties may split the Dotlet diagonals into two alignments. Lower gap-opening or gap-extension penalties, on the other hand, can cause diagonals to fuse. What happens here depends on the length of the gaps that Lalign needs to insert in the local alignments.

When it comes to judging the quality of your alignment, size definitely matters. It isn't uncommon to see short segments with very good scores. This can result from a variety of reasons — such as low-complexity segments, small-scale regularity in protein sequences, recurrent patterns of hydrophobic/hydrophilic residues, and so on. For instance, in Figure 8-11, the second and the third local alignments are obviously meaningless because they are too short. However, their E-values are better than that of the top alignment. This clearly illustrates that a short alignment can easily get a good score by chance.

Bioinformatics programs aren't perfect: They can make mistakes and deliver wrong results. Unfortunately, right and wrong obey the same practical rules in bioinformatics as they do in life: They are not absolute concepts.

The top alignment that Lalign reports here illustrates this limitation well. Because of the dot-plot information, we trust this alignment to be correct overall. On the other hand, the two sequences are so remotely related — and have so little similarity — that we must be cautious in our interpretation: Many amino acids could be incorrectly aligned, which means this alignment is probably right *and* wrong. The only way to really push this analysis further would be to make a multiple sequence alignment or to use structural information. (For more on multiple sequence alignments, see Chapter 9; for more on structures, see Chapter 11.)

Making Global Alignments over the Internet

A global alignment is just what the name implies: an alignment of every amino acid or nucleotide found in your sequences. When you decide to make a global alignment, you implicitly make the hypothesis that your two sequences are related over their entire lengths.

Global alignments are very important for making multiple sequence alignments. However, when done using only two sequences, they offer little interest for biological analysis. Whenever you have two closely related sequences, you can generate a much more informative result by gathering a few more related sequences and making a multiple sequence alignment. (Chapter 9 tells you everything you need to know for this purpose.)

Global alignments aren't useful at all for discovering similarities between two sequences because the statistical method for evaluating E-values doesn't apply to them. Even so, beginners often prefer global alignments to local because they're simpler to understand.

In a global alignment, no amino acid or nucleotide mysteriously disappears: You find in the output exactly what you put in the box, plus a few extra gaps. There are three main reasons for making a global alignment:

- **Checking minor differences between two sequences.** This may happen with data that you've manipulated and possibly altered. The global alignment is the best way to localize potential problems.
- **Analyzing polymorphisms (for example, SNPs) between closely related sequences.**
- **Comparing two sequences that partly overlap.** In that case, you want to make a global pairwise comparison that doesn't penalize misalignments at the extremities of the sequences.

Using Lalign to Make a Global Alignment

In terms of algorithms, there is not much difference between a local and a global alignment. If you were to write the programs for both, you would not find more than one line of difference between these two algorithms!

You can use Lalign to produce global alignments, but there are other choices — some of which we've listed in Table 8-4, at the end of this chapter.

Using Lalign to produce a global alignment is exactly like using Lalign to produce a local alignment. The only thing extra you need to do is to is to click the radio button that says *Global without End-Gap Penalty,* near the top of the Lalign form at

```
ww.ch.embnet.org/software/LALIGN_form.html
```

Aligning Proteins and DNA

The alignment tools we describe above are only truly effective when you want to compare sequences of a similar type — proteins with proteins or DNA with DNA. Sometimes, however, you need to compare a protein with a piece of DNA (its original gene, for instance). Doing such an alignment requires special tools that can insert the long gaps corresponding to the introns. We know at least two places where you can do this online. One is at the Pasteur Institute (`bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html`) and the other is maintained by Dr. Peer Bork and his group at the European Molecular Biology Laboratory in Heidelberg (`coot.embl.de/pal2nal`). Both of these servers will require you to have your protein and the corresponding DNA sequence ready before you can use them.

If you only have the protein, you can either run a blastx against a complete genome (see Chapter 6) or use a Web server called Protogene (available at `www.tcoffee.org`). Protogene automatically fetches the DNA sequence corresponding to a given protein.

Free Pairwise Sequence Comparisons over the Internet

Many existing servers propose pairwise comparisons. Table 8-4 lists a brief selection of some of the more stable resources.

Table 8-4 Online Pairwise Alignment Programs

Name	*Address*	*Alignment Type*
Lalign	`www.ch.embnet.org/software/LALIGN_form.html`	Global/Local
Lalign	`fasta.bioch.virginia.edu/fasta_www/plalign.htm`	Global/Local
USC	`www-hto.usc.edu/software/seqaln/seqaln-query.html`	Global/Local/
Alion	`fold.stanford.edu/alion/`	Global/Local
Align	`genome.cs.mtu.edu/align.html`	Global/Local
Align	`www.ebi.ac.uk/emboss/align/`	Global/Local
xenAliTwo	`www.soe.ucsc.edu/~kent/xenoAli/xenAliTwo.html`	Local for DNA
Blast2seqs	`www.ncbi.nlm.nih.gov/blast/bl2seq/wblast2.cgi`	Local BLAST
Protal2dna	`bioweb.pasteur.fr/seqanal/interfaces/protal2dna.html`	Protein against DNA
Pal2nal	`coot.embl.de/pal2nal`	Protein against DNA

REMEMBER

Generating an alignment isn't always enough. You may also need to visualize this alignment and to evaluate its statistical significance. Table 8-5 lists a few sites that enable you to do this online.

Table 8-5 Online Pairwise Alignments Analyses

Name	*Address*	*Function*
lalnview	`www.expasy.ch/tools/sim-prot.html`	Visualization
prss	`www.ch.embnet.org/software/PRSS_form.html`	Evaluation
prss	`Fasta.bioch.virginia.edu/fasta/prss.htm`	Evaluation
graph-align	`darwin.nmsu.edu/cgi-bin/graph_align.cgi`	Evaluation

Chapter 9

Building a Multiple Sequence Alignment

In This Chapter

- Gathering the sequences you need to make a multiple sequence alignment
- Making a multiple sequence alignment with ClustalW and MUSCLE
- Creating and comparing multiple sequence alignments with Tcoffee
- Comparing sequences you cannot align

> *The man with two watches never knows the time, and the man with one watch only thinks he knows.*
>
> — A man with multiple watches

In this chapter, we show you how to compare many protein or nucleotide sequences simultaneously. This is something you probably want to do if you have a sequence and you want to elucidate the role of each amino acid or nucleotide that it contains.

To compare multiple sequences, your first job is to select a few sequences and then align them all. In this chapter, we show you how to select the most appropriate sequences and how you can generate such a *multiple sequence alignment.* We introduce three different multiple-sequence-alignment methods that should cover most of your needs: ClustalW because everybody uses it, MUSCLE because it is very fast, Tcoffee because it is very accurate, and can let you combine sequences and structures. Since it's too easy to generate bad alignments that look good, we also show you how to evaluate the quality of your alignment.

Sometimes your multiple sequence alignment isn't good enough to be useful. When this happens, we show you alternatives to sequence alignment — and that it's possible to compare your sequences without aligning them. We introduce two of these methods: the Gibbs sampler (for identifying related segments of the same length) and Pratt, a motif-discovery method for discovering PROSITE patterns.

A multiple alignment is probably the most useful investigative tool you can generate in bioinformatics. In many ways, multiple sequence alignments are to bioinformatics what Swiss knives are to MacGyver; we're hard-pressed to think of a situation where they don't help. If you aren't sure whether a multiple alignment can help your work — or if you don't see exactly how you can fit one of these into your experiments — we suggest that you start reading the first section of this chapter, "Finding Out if a Multiple Sequence Alignment Can Help You."

In any case, remember that multiple alignments are useful for predicting protein structures (see Chapter 11), central for predicting the function of proteins, and indispensable for phylogenetic analysis (see Chapter 13). Of course, the better your multiple sequence alignment, the better use you can make of it: better structural models, better functional predictions, and better phylogenetic trees.

Even so, building multiple sequence alignments is far from an exact science. In fact, it's more art than science, requiring that you use everything you know in bioinformatics and in biology. This chapter is here to give you the secrets of the trade so that you, too, will be able to create the multiple sequence alignment that best suits your needs — and even find your own recipes.

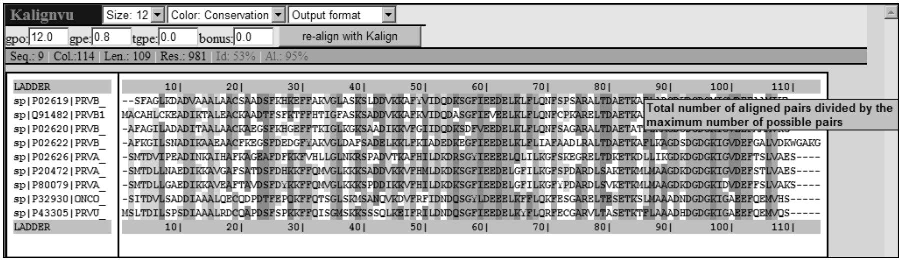
If you already have an alignment and you only want to modify it or improve its appearance for publication (highlighting, shading, and so on), go directly to Chapter 10 — that's the chapter entirely dedicated to such fine-tuning techniques.

Finding Out if a Multiple Sequence Alignment Can Help You

Figure 9-1 shows you what a multiple alignment actually looks like: As you can see, it's really only a matter of rewriting your sequences so similar features end up in the same columns. That's it, plain and simple.

REMEMBER

This kind of multiple alignment is ideal if you want to study a sequence family where all sequences share the *same common ancestor.* Don't worry if you have only one member at this stage of the game; we show you later in this chapter how to organize a family gathering!



gpo: 12.0 gpe: 0.8 tgpe: 0.0 bonus: 0.0

Seq.: 9 | Col.:114 | Len.: 109 | Res.: 981 | Id: 53% | Al.: 95%

```
LADDER                    10|       20|       30|       40|       50|       60|       70|       80|       90|      100|      110|
sp|P02619|PRVB_   --SFAGLKDADVAAALAACSAADSFKHKEFFAKVGLASKSLDDVKKAFYVIDQDKSGFIEEDELKLFLQNFSPSARALTDAETKA
sp|Q91482|PRVB1   MACAHLCKEADIKTALEACKAADTFSFKTFFHTIGFASKSADDVKKAFKVIDQDASGFIEVEELKLFLQNFCPKARELTDAETKA
sp|P02620|PRVB_   -AFAGILADADITAALAACKAEGSFKHGEFFTKIGLKGKSAADIKKVFGIIDQDKSDFVEEDELKLFLQNFSAGARALTDAETAT
sp|P02622|PRVB_   -AFKGILSNADIKAAEAACFKEGSFDEDGFYAKVGLDAFSADELKKLFKIADEDKEGFIEEDELKLFLIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG
sp|P02626|PRVA_   -SMTDVIPEADINKAIHAFKAGEAFDFKKFVHLLGLNKRSPADVTKAFHILDKDRSGYIEEEELQLILKGFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAES----
sp|P20472|PRVA_   -SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDKSGFIEEDELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----
sp|P80079|PRVA_   -SMTDLLGAEDIKKAVEAFTAVDSFDYKKFFQMVGLKKKSPDDIKKVFHILDKDKSGFIEEDELGFILKGFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----
sp|P32930|ONCO_   -SITDVLSADDIAAALQECQDPDTFEPQKFFQTSGLSKMSANQVKDVFRFIDNDQSGYLDEEELKFFLQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----
sp|P43305|PRVU_   MSLTDILSPSDIAAALRDCQAPDSFSPKKFFQISGMSKKSSSQLKEIFRILDNDQSGFIEEDELKYFLQRFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----
LADDER                    10|       20|       30|       40|       50|       60|       70|       80|       90|      100|      110|
```

Total number of aligned pairs divided by the maximum number of possible pairs

Figure 9-1: A multiple sequence alignment.

Identifying situations where multiple alignments do not help

Part of knowing when to use multiple sequence alignments involves knowing when *not* to use them. The multiple-sequence methods we describe in this chapter don't work well for assembling the sequence pieces in a sequencing project. If you have a set of *shotgun sequences* — short, partly overlapping sequences — or if you want to turn an EST cluster into a gene sequence, multiple sequence alignment doesn't work well — and dumping your sequences into ClustalW, MUSCLE, or Tcoffee can produce a very disappointing and uninformative result.

If you're facing this particular type of problem, you may want to use specialized sequence assembly tools, such as Phred and Phrap (`www.phrap.org`) or cap3 (`www.mainlab.clemson.edu/cgi-bin/gdr/gdr_cap3`). Unfortunately, the intricacies of specialized sequence-assembly tools (which we briefly overview in Chapter 5) are far beyond the scope of this book.

Another situation that *doesn't* call for multiple sequence alignment is when the sequence you're interested in has no homologue in any of the sequence databases. If this happens, you're out of luck; there's no way you can build a multiple alignment that relates to this sequence. You may try to find a few sequences by using functional criteria and conducting a pattern search with Pratt, but don't count on any miracles.

Helping your research with multiple sequence alignments

With the appropriate sequences in hand, you can build the proper multiple sequence alignment. To be honest, you could build this alignment manually, rather than by using a computer. Manual alignments are like home-cooked food: When prepared by a good cook, nothing beats them; for regular folks like us, however, the automatic processing (TV dinner) is generally much better.

The idea behind a multiple alignment like the one in Figure 9-1 is to put amino acids or nucleotides in the same column because they're similar according to some criterion. You can use four major criteria to build a multiple alignment of sequences that all have different properties, as Table 9-1 conveniently shows.

Table 9-1 Main Criteria for Building a Multiple Sequence Alignment

Criterion	*Meaning*
Structural similarity	Amino acids that play the same role in each structure are in the same column. Structure-superposition programs are the only ones that use this criterion.
Evolutionary similarity	Amino acids or nucleo tides related to the same amino acid (or nucleotide) in the common ancestor of all the sequences are put in the same column. No automatic program explicitly uses this criterion, but they all try to deliver an alignment that respects it.
Functional similarity	Amino acids or nucleotides with the same function are in the same column. No automatic program explicitly uses this criterion, but if the information is available, you can force some programs to respect it — or you can edit your alignment manually.
Sequence similarity	Amino acids in the same column are those that yield an alignment with maximum similarity. Most programs use sequence similarity because it is the easiest criterion. When the sequences are closely related, their structural, evolutionary, and functional similarities are equivalent to sequence similarity.

The first three criteria have a clear biological meaning — the fourth one doesn't. Yet, when the sequences are similar enough, you can use similarity to produce a multiple alignment that reflects the evolutionary, the structural, and the functional relationships that exist between your sequences. Of course, to do this, you need an alignment that you can trust. Later in this chapter, we show you how to make sure you're putting your money on the right alignment.

Table 9-2 lists the main applications of multiple sequence alignments.

Table 9-2 Main Applications of Multiple Sequence Alignments

Application	*Procedure*
Extrapolation	A good multiple alignment can help convince you that an uncharacterized sequence is really a member of a protein family. Alignments that include Swiss-Prot sequences are the most informative. Use the ExPASyBLAST server (at `www.expasy.ch/tools/blast/`) to gather and align them.
Phylogenetic analysis	If you carefully choose the sequences you include in your multiple alignment, you can reconstruct the history of these proteins. Use the Pasteur Phylip server at `bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`.
Pattern identification	By discovering very conserved positions, you can identify a region that is characteristic of a function (in proteins or in nucleic-acid sequences). Use the logo server for that purpose: `www-lmmb.ncifcrf.gov/~toms/sequencelogo.html`.
Domain identification	It is possible to turn a multiple sequence alignment into a profile that describes a protein family or a protein domain (PSSM). You can use this profile to scan databases for new members of the family. Use NCBI-BLAST to produce and analyze PSSMs: `www.ncbi.nlm.nih.gov/blast/blastcgihelp.shtml#pssm`.
DNA regulatory elements	You can turn a DNA multiple alignment of a binding site into a weight matrix and scan other DNA sequences for potentially similar binding sites. Use the Gibbs sampler to identify these sites: `bayesweb.wadsworth.org/gibbs/gibbs.html`
Structure prediction	A good multiple alignment can give you an almost perfect prediction of your protein secondary structure for both proteins and RNA. Sometimes it can also help in the building of a 3-D model.
nsSNP analysis	Various gene alleles often have different amino-acid sequences. Multiple alignments can help you predict whether a Non-Synonymous Single-Nucleotide Polymorphism is likely to be harmful. See the SIFT site for more details: `blocks.fhcrc.org/sift/SIFT.html`.
PCR analysis	A good multiple alignment can help you identify the less-degenerated portions of a protein family, in order to fish out new members by PCR (polymerase chain reaction). If this is what you want to do, you can use the following site: `blocks.fhcrc.org/codehop.html`.

Whatever your choice or motivation, always bear in mind that the reason you can use multiple alignments for so many things is that they can give you an instant picture of the forces that shape evolution! When you look at proteins or DNA sequences, the rules are ruthless:

- Important amino acids (or nucleotides) are not allowed to mutate. For instance, active sites of enzymes are much conserved.
- Less-important residues change more easily — sometimes randomly — and sometimes in order to adapt a function.

For you, this has a very simple consequence: When you look at a multiple alignment, you can make the hypothesis that *conserved* positions (columns where all the sequences contain the same amino acid or nucleotide) are more important for the function than *non conserved* positions (columns where the sequences contain different amino acids or nucleotides). Of course, you could tell all this from the alignment of just two sequences — but using more sequences makes it easier to discriminate between important and less-important positions.

And one more thing: A multiple sequence alignment is a terrific way to present your results. It lets you put lots of information into a single model and makes it incredibly easy to spot inconsistencies or potential problems. If you want to get an important point across — in a paper or in front of an audience — a good, hand-crafted, well-colored multiple alignment says more than 3 billion nucleotides in bulk.

All this being said, multiple alignments do NOT cure asthma, arthritis, lumbago, or baldness. They won't make you stronger or increase any of your potentials (except maybe your research potential). But at this point, if you need more arguments to convince you of the genius of multiple alignments, chances are you don't really need to make this kind of alignment.

Choosing the Right Sequences

Anyone who ever worked in a lab knows that molecular biology is very much like cooking: It's all about selecting the right ingredients and putting them into the pot at the right time and in the right order.

Building a multiple alignment obeys the same rule. Before you build your alignment, you must carefully select the sequences you want to align. These sequences are members of the same protein family, and they all share a common ancestor. The family is usually too large to be entirely included in your multiple alignment, and picking the right sequences is an art. If you want to be good at this game, you need to know what you want to show with your alignment — and you need to know how the multiple alignment programs work.

Of course, to understand these programs well, you need to see them in action — and for that, you need to have a few sequences! This is a typical chicken/egg problem for which no simple solution exists. We suggest that you go through the next section once, and come back again after you start playing with the multiple sequence alignments that we describe in the other parts of the chapter. After one or two rounds, things should click into place.

The kinds of sequences you're looking for

Let's assume that you start this procedure with your favorite sequence. You want to thoroughly study this sequence and you know that in order to do this you have to build a multiple alignment. Table 9-3 summarizes most of the things you must take into account when selecting these extra sequences.

Table 9-3 A Few Guidelines for Selecting Sequences

Problem	***Diagnostic***
Proteins or DNA	Use proteins whenever possible. You can turn them back into DNA *after* doing the multiple alignment.
Many sequences	Start with 10–15 sequences; avoid aligning more than 50 sequences.
Very different sequences	Sequences that are less than 30 percent identical to more than half the other sequences in the set often cause troubles.
Identical sequences	They never help. Unless you have a *very* good reason to do so, avoid incorporating into your multiple alignment any sequence that's more than 90 percent identical to another sequence in the set.
Partial sequences	Multiple-sequence-alignment programs prefer sequences that are roughly the same length. Programs often have difficulties comparing items in a mixture of complete sequences and shorter fragments.
Repeated domains	Sequences with repeated domains cause trouble for most multiple-alignment programs — especially if the number of domains is different. When this happens, you may be better off extracting the domains yourself with Dotlet or Lalign (see Chapter 8) and making a multiple alignment of those segments.

Using DNA or protein sequences

If the sequences you're interested in are non-coding sequences, you obviously have no choice — you must use DNA. However, beware that non-coding DNA sequences can be tricky to align. If you cannot generate a proper alignment from sequences that you know are related, you could use a local multiple alignment method, such as the Gibbs sampler, or a pattern extraction motif, such as Pratt. (See "Comparing Sequences That You Can't Align," later in this chapter, for a description of these two methods.)

Multiple-sequence-alignment methods are at their best when aligning protein sequences. The reason is that protein sequences are three times shorter than the corresponding DNA, and they use a more informative alphabet of 20 amino acids.

If you want to persist in carrying out a phylogenetic analysis on a set of coding DNA sequences, things work better if you do the following:

1. Translate your DNA sequences into proteins.
2. Perform the multiple alignments on the proteins.
3. Thread the DNA back onto the protein multiple sequence alignment framework using pal2nal (`coot.embl.de/pal2nal`) or Protogene if you do not have the original DNA sequence (`www.tcoffee.org`).

If your proteins are difficult to align because they have few similarities, DNA information does not help. This has to do with the degeneracy of the genetic code and the fact that there are 20 amino acids and only 4 nucleotides. If there is little signal at the protein level, you can be sure that there is NO useful signal at the DNA level.

Choosing the right number of sequences

Of course, there's no absolute answer to this question, such as *42* or *7*. A few years ago, the answer was easy: Get everything you can and go to the lab if there aren't enough sequences in the databases! But that isn't true anymore. These days — given the sizes of the databases and new complete genome sequences flowing in twice a month — you may easily find hundreds or thousands of sequences that would be suitable for inclusion in your multiple sequence alignment. But that doesn't mean you have to use them all!

In our opinion, you should start with a relatively small number of sequences — between 10 and 15 sequences would be suitable for most cases. After you get something interesting happening with this small set, you can always increase its size. In any case, it's hard to see any reason for generating a multiple alignment with more than 50 sequences, unless you're interested in building some extensive phylogenetic tree.

If you start with hundreds of sequences, you immediately hit troubles. There are good reasons why:

- ***Computing* big alignments is difficult.** Public servers do not have infinite resources. Your job may take a very long time to run (if it runs). For you, this makes it difficult to tune parameters and check alternatives.
- ***Building* big alignments is difficult.** Multiple alignment programs are not very good at handling very large sets of sequences.
- ***Displaying* big alignments is difficult.** You can't print them, and they clog your computer when you want to visualize them. If columns are longer than one page, interpretation becomes impossible.
- ***Using* big alignments is difficult.** Tree-building and structure-prediction programs cannot handle them easily.
- **Making *accurate* big alignments is difficult.** You want your multiple sequence alignment to be highly reliable so you can be confident that all the sequences it contains are true members of the family. A major cause for concern is that multiple-sequence-alignment programs make mistakes. The curse is that these mistakes do not add up, they *multiply* — making it easy for a tiny number of bad sequences to ruin an entire alignment. Of course, the more sequences you have, the more likely this is to happen. The best way to avoid such a disaster is to start small — and gradually increase the size of your multiple sequence alignment until it contains all the sequences you're interested in.

Now that you know how many sequences you need, the last question you face is deciding how related these sequences must be. Should you choose sequences that are very similar or very different?

Making the right compromise between similarity and new information

If you think that very similar sequences give very good alignments, you're right! However, a multiple sequence alignment that's correct isn't enough; it must also be useful.

REMEMBER

For instance, an alignment that only contains very similar sequences brings little information. You can use it to extrapolate annotations, but you can't do phylogeny, structure prediction, or any of the other useful applications that we list in Table 9-2. These other tasks require being able to observe mutation patterns in every column — which isn't possible if you have an alignment in which most columns are entirely conserved.

Grabbing the most distantly related sequences you can find doesn't work, either. Multiple-sequence-alignment programs can't use sequences that are too different — even if these sequences are homologous. In fact, two things multiple-sequence-alignment programs *really* don't like are

- Sequences that are very different from every other sequence in the group
- Sequences that need long insertions/deletions to be properly aligned

Selecting your sequences for a multiple alignment is a team-building exercise: You want everyone to be a bit special, but you don't want your team members to be so different that they can't communicate. Teams of divas rarely go very far — and neither do gangs of sextuplets.

When you select your sequences, the general rule is that you want them to be as distantly related as possible — without requiring too many gaps in order to be properly aligned. These two criteria are mutually exclusive, so finding the right trade-off requires a bit of strategy. The following steps show you a general approach that you should have in mind when gathering your sequences:

1. **Select a few sequences.**

   See the section "Gathering your sequences with online BLAST servers," later in this chapter.

2. **Compute a multiple alignment by using one of the servers we introduce in this chapter.**

   See the section "Choosing the Right Method of Multiple Sequence Alignment," later in this chapter.

3. **Evaluate the quality of your alignment visually.**

   See the section "Interpreting Your Multiple Sequence Alignment," later in this chapter.

4. **If your alignment looks good, keep the sequences.**

   A good alignment contains nicely conserved blocks separated by regions with insertions and deletions. If you have such an alignment, your sequence set is probably appropriate — and you can try to extend it by adding a few new sequences.

5. **If your alignment is difficult to interpret,**

   a. Examine the sequences more closely — try to remove the troublemakers that are the most distantly related, or those that cause long insertions/deletions.

   b. Redo the alignment with the smaller set.

   c. Keep trimming the set until you get a multiple alignment that's easier to interpret.

TIP

If you can, make sure that *each sequence is between 30 and 70 percent identical with more than half of the sequences in the set.* This way, you're making a reasonable trade-off between new information and alignment quality.

TIP

Before adding a sequence to a multiple alignment, you can try to figure out whether it's a good choice by making pairwise comparisons with some of the tools we describe in Chapter 6. (*Pairwise* comparisons let you compare sequences two by two.) However, we don't recommend that you do so exhaustively; it's time-consuming.

Naming your sequences the right way

Naming may sound like a trivial issue, but it's not. Multiple-sequence-alignment programs have no standard way of handling the sequence names. If you stick to the following four rules, however, you will NEVER get into trouble (if you don't, you're on your own):

- Never use white spaces in your sequence names.
- Do not use special symbols. Stick to plain letters, numbers, and the underscore (_) to replace spaces. Avoid ALL other symbols, especially those that are the most tempting for special sequences (such as @, #, _,^ and so on).
- Never use names longer than 15 characters.
- Never give the same name to two different sequences in your set. Although some programs accept it, others (such as ClustalW) don't.

If you don't obey these naming rules, some multiple-sequence-alignment programs may automatically change the name of your sequences, without the courtesy of telling you.

Gathering your sequences with online BLAST servers

There are two types of sequences that you may want to integrate into your multiple alignment:

- **Characterized sequences:** These are sequences for which you have good annotations and experimental information. You'd definitely want to include these sequences in your alignment because they bring biological information with them — and also allow feature propagation.
- **Uncharacterized sequences:** This category can include your sequence(s) of interest as well as database sequences. Uncharacterized sequences must be members of the same family. Your main motivation in including them in your multiple alignment is to distinguish between the conserved positions that cannot mutate and the other, less-important columns. They help in getting some contrast on your sequence of interest.

In this section, we show you how to gather these sequences with the BLAST database search program. If you want to know all the gory details about BLAST, check out Chapter 7. Here we give you only the bare minimum to get by.

With BLAST, you can search databases for sequences that are homologous (similar) to a query. The original query can be any sequence you are interested in — protein or DNA — and you can use BLAST to search both protein and DNA databases. The main reason for using BLAST is to identify database sequences that are so similar to the query that they probably are homologous. We commonly refer to such sequences as *hits* or *matches.*

There are many BLAST servers around — as becomes abundantly clear in Chapter 7 — and you're free to try any (or all) of them. Nonetheless, they aren't all equally convenient for retrieving the sequences you're interested in. Table 9-4 lists three BLAST servers that are useful for generating a list of sequences in FASTA format (the best format for moving your sequences around, from one program to another) or for sending chosen sequences to a multiple-alignment server.

Table 9-4 BLAST Servers Integrating Multiple-Alignment Methods

Address	*What You Can Do There*
`www.expasy.ch/tools/blast/`	Extract entire sequences, export sequences in FASTA, submit sequences to ClustalW, Tcoffee or MAFFT. Turn the list of Hits into a non-redundant collection of sequences.
`npsa-pbil.ibcp.fr/cgi-bin/npsa_automat.pl?page=npsa_blast.html`	Extract entire sequences, extract sequences fragments, export sequences in FASTA, submit sequences to ClustalW.
`srs.ebi.ac.uk`	Submit sequences to ClustalW.

In the following steps list, we show you how to use the ExPASy server. The SRS server, although sometimes useful, is a bit harder to use than the other two listed in Table 9-4.

Selecting sequences on the ExPASy server

In the following steps list, we select appropriate sequences to make a multiple sequence alignment of calcium-dependent kinase proteins.

You can use this server only to retrieve *protein* sequences. If you're interested in gathering *DNA* sequences, use the European Bioinformatics SRS server (`srs.ebi.ac.uk`) instead.

The BLAST server of the PBIL site is very similar to ExPASy. If you don't find the database you're interested in on the ExPASy server, try the PBIL.

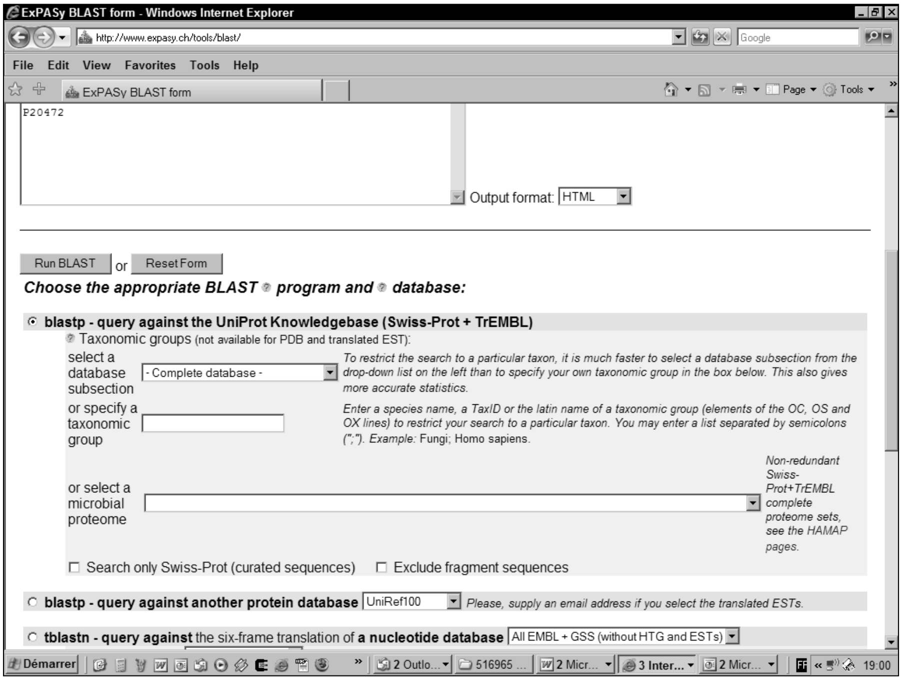
1. **Point your browser to** `www.expasy.ch/tools/blast/`**.**

   The BLAST page of the ExPASy server appears.

2. **Enter the Sequence Accession Number** P20472**, as shown in Figure 9-2).**

   This is the accession number of the human parvalbumin.

   If you prefer, you can also paste the sequence in raw format (that is, use the sequence only, without any header). The program ignores spaces and numbers.


ExPASy BLAST form - Windows Internet Explorer
http://www.expasy.ch/tools/blast/
Google
File Edit View Favorites Tools Help
ExPASy BLAST form
Page Tools
Run BLAST
Reset Form
Démarrer
2 Outlo...
516965 ...
2 Micr...
3 Inter...
2 Micr...
19:00

```
P20472
```

Output format: HTML

Run BLAST or Reset Form

Choose the appropriate BLAST program and database:

● **blastp - query against the UniProt Knowledgebase (Swiss-Prot + TrEMBL)**

Taxonomic groups (not available for PDB and translated EST):

select a database subsection: - Complete database -

To restrict the search to a particular taxon, it is much faster to select a database subsection from the drop-down list on the left than to specify your own taxonomic group in the box below. This also gives more accurate statistics.

or specify a taxonomic group

Enter a species name, a TaxID or the latin name of a taxonomic group (elements of the OC, OS and OX lines) to restrict your search to a particular taxon. You may enter a list separated by semicolons (";"). Example: Fungi; Homo sapiens.

or select a microbial proteome

Non-redundant Swiss-Prot+TrEMBL complete proteome sets, see the HAMAP pages.

□ Search only Swiss-Prot (curated sequences) □ Exclude fragment sequences

○ **blastp - query against another protein database** UniRef100 *Please, supply an email address if you select the translated ESTs.*

○ **tblastn - query against** the six-frame translation of **a nucleotide database** All EMBL + GSS (without HTG and ESTs)

Figure 9-2: Using the BLAST ExPASy server to gather sequences.

3. **Select the BLAST flavor that you're interested in.**

   If you gave a protein sequence in Step 2, select blastp.

   If you gave a coding DNA sequence in Step 2, select tblastn.

4. **Keep the default option — Complete Database — in the pull-down menu.**

   This amounts to simultaneously searching Swiss-Prot + TrEMBL + TrEMBL_NEW. If the search reports too many sequences that are very similar to your sequence of interest, you can decrease the number of identical hits by selecting a smaller database from the Database pull-down menu — Swiss-Prot, for example, or the database of only the microbial proteomes.

   WARNING!

   For a multiple sequence alignment, do not select Translated ESTs from the pull-down menu. These sequences are mostly incomplete protein sequences that may confuse the multiple-alignment procedure.

5. **Scroll down to the Options section and set the Number of Best Scoring Sequences to Show option to 1000.**

   Doing this makes it more likely that you'll find appropriate sequences in the BLAST result for your multiple alignment.

6. **In the same Options section, set the Number of Best Alignments to Show option to 1000.**

   This choice makes it possible to judge the quality of the alignment before selecting a sequence.

7. **Click the Run BLAST button.**

   After a brief pause, a Results page appears.

8. **Scroll down the page to select the sequences you want.**

   You select a sequence by checking the box to its left.

   This is the most delicate part of the process. There is no absolute rule to selecting your sequences, but you can use the following guidelines:

   - **Select the top sequence.** This top sequence is usually your sequence of interest. If your sequence of interest is not at the top, you may have to add it to the list later on.
   - **For a first analysis, you want to select ten sequences or fewer.** Ideally, the ten sequences to select should be evenly spaced between the very good E-values (10-^{40}) and less-good E-values (10-^{5}). Figure 9-3 gives you an idea of what such a selection looks like.

     For the purpose of this steps list, choose (from top to bottom): P20472, P80079, P02626, P02619, P43305, P32930, Q91482, P02620, P02622, P02627.
   - **Before selecting a sequence, check to make sure it's similar to the query sequence — along its entire length.**

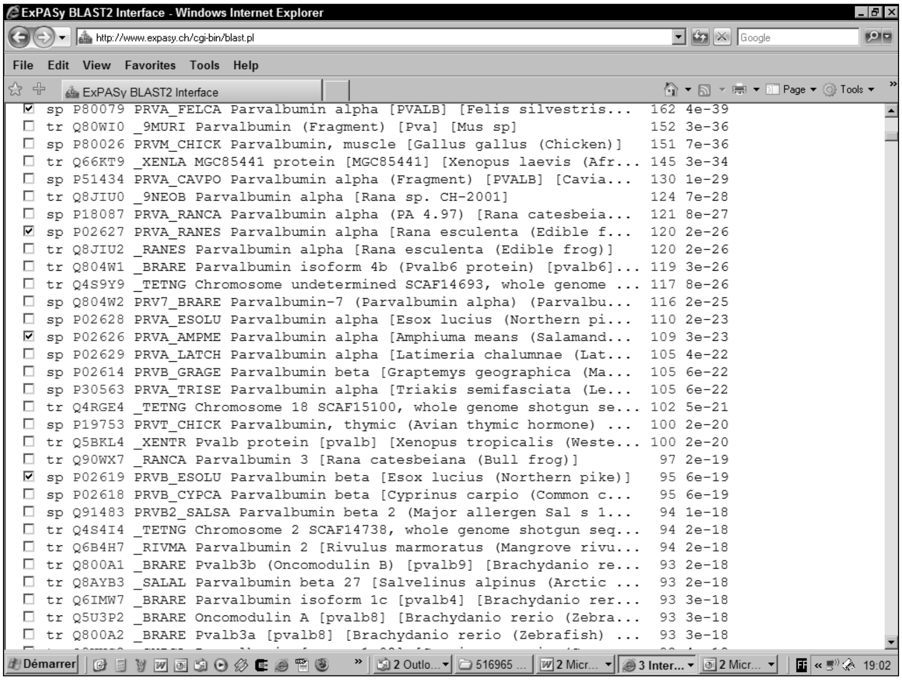
     The alignment section is at the bottom of the BLAST output. You must be especially careful with hits that have E-values higher than 10-^{10}. They are equally likely to correspond to a good partial match, a global overall match, or a match between a protein fragment and your sequence. Inspecting the alignment is the only way to distinguish between these situations.

9. **Choose the method you want to use to export your sequences from the Send Selected Sequences pull-down menu, as shown in Figure 9-4.**

   There are several ways to export your sequences:

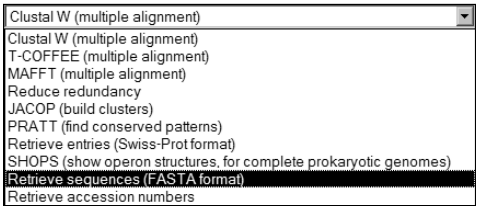
   - **FASTA:** Generates a file that contains your sequences in FASTA format. You can save this file with the File⇨Save As option of your browser. When you need to, you can reopen this file with your browser, in order to cut and paste its content into another server form (MUSCLE for instance, at `www.drive5.com/muscle/`).

- **ClustalW, Tcoffee, and MAFFT:** These are multiple-sequence-alignment packages running on the EMBnet server. Select any of these to align the selected sequences.
- **Reduce Redundancy:** This option will extract the most meaningful sequences from your dataset. Ideal if you have too many sequences and you don't know how to choose.
- **Pratt:** Will search for conserved motifs in your sequences without aligning them.



```
☑ sp P80079 PRVA_FELCA Parvalbumin alpha [PVALB] [Felis silvestris...  162 4e-39
☐ tr Q80WI0 _9MURI Parvalbumin (Fragment) [Pva] [Mus sp]              152 3e-36
☐ sp P80026 PRVM_CHICK Parvalbumin, muscle [Gallus gallus (Chicken)]   151 7e-36
☐ tr Q66KT9 _XENLA MGC85441 protein [MGC85441] [Xenopus laevis (Afr... 145 3e-34
☐ sp P51434 PRVA_CAVPO Parvalbumin alpha (Fragment) [PVALB] [Cavia...  130 1e-29
☐ tr Q8JIU0 _9NEOB Parvalbumin alpha [Rana sp. CH-2001]                124 7e-28
☐ sp P18087 PRVA_RANCA Parvalbumin alpha (PA 4.97) [Rana catesbeia...  121 8e-27
☑ sp P02627 PRVA_RANES Parvalbumin alpha [Rana esculenta (Edible f...  120 2e-26
☐ tr Q8JIU2 _RANES Parvalbumin alpha [Rana esculenta (Edible frog)]    120 2e-26
☐ tr Q804W1 _BRARE Parvalbumin isoform 4b (Pvalb6 protein) [pvalb6]... 119 3e-26
☐ tr Q4S9Y9 _TETNG Chromosome undetermined SCAF14693, whole genome ... 117 8e-26
☐ sp Q804W2 PRV7_BRARE Parvalbumin-7 (Parvalbumin alpha) (Parvalbu...  116 2e-25
☐ sp P02628 PRVA_ESOLU Parvalbumin alpha [Esox lucius (Northern pi...  110 2e-23
☑ sp P02626 PRVA_AMPME Parvalbumin alpha [Amphiuma means (Salamand...  109 3e-23
☐ sp P02629 PRVA_LATCH Parvalbumin alpha [Latimeria chalumnae (Lat...  105 4e-22
☐ sp P02614 PRVB_GRAGE Parvalbumin beta [Graptemys geographica (Ma...  105 6e-22
☐ sp P30563 PRVA_TRISE Parvalbumin alpha [Triakis semifasciata (Le...  105 6e-22
☐ tr Q4RGE4 _TETNG Chromosome 18 SCAF15100, whole genome shotgun se... 102 5e-21
☐ sp P19753 PRVT_CHICK Parvalbumin, thymic (Avian thymic hormone) ...  100 2e-20
☐ tr Q5BKL4 _XENTR Pvalb protein [pvalb] [Xenopus tropicalis (Weste... 100 2e-20
☐ tr Q90WX7 _RANCA Parvalbumin 3 [Rana catesbeiana (Bull frog)]         97 2e-19
☑ sp P02619 PRVB_ESOLU Parvalbumin beta [Esox lucius (Northern pike)]   95 6e-19
☐ sp P02618 PRVB_CYPCA Parvalbumin beta [Cyprinus carpio (Common c...   95 6e-19
☐ sp Q91483 PRVB2_SALSA Parvalbumin beta 2 (Major allergen Sal s 1...   94 1e-18
☐ tr Q4S4I4 _TETNG Chromosome 2 SCAF14738, whole genome shotgun seq...  94 2e-18
☐ tr Q6B4H7 _RIVMA Parvalbumin 2 [Rivulus marmoratus (Mangrove rivu...  94 2e-18
☐ tr Q800A1 _BRARE Pvalb3b (Oncomodulin B) [pvalb9] [Brachydanio re...  93 2e-18
☐ tr Q8AYB3 _SALAL Parvalbumin beta 27 [Salvelinus alpinus (Arctic ...  93 2e-18
☐ tr Q6IMW7 _BRARE Parvalbumin isoform 1c [pvalb4] [Brachydanio rer...  93 3e-18
☐ tr Q5U3P2 _BRARE Oncomodulin A [pvalb8] [Brachydanio rerio (Zebra...  93 3e-18
☐ tr Q800A2 _BRARE Pvalb3a [pvalb8] [Brachydanio rerio (Zebrafish) ...  93 3e-18
```

Figure 9-3: Selecting sequences on the BLAST output.



Clustal W (multiple alignment)

- Clustal W (multiple alignment)
- T-COFFEE (multiple alignment)
- MAFFT (multiple alignment)
- Reduce redundancy
- JACOP (build clusters)
- PRATT (find conserved patterns)
- Retrieve entries (Swiss-Prot format)
- SHOPS (show operon structures, for complete prokaryotic genomes)
- Retrieve sequences (FASTA format)
- Retrieve accession numbers

Figure 9-4: Output selector on the ExPASy BLAST page.

Gathering a known collection of sequences from Swiss-Prot

If you already know the name or accession number of every sequence you want to include in your multiple alignment and if these sequences are in Swiss-Prot or in TrEMBL, you can directly access them by using a special online ExPASy facility.

This can be handy if you're using a BLAST server that does not have sequence-extraction facilities. You can simply copy the accession number of the sequences you are interested in, and extract them here.

WARNING!

This procedure works only if the sequences that you're interested in belong to the Swiss-Prot or TrEMBL databases. The following steps list shows you how to use this server:

1. **Point your browser to** `www.expasy.ch/sprot/sprot-retrieve-list.html`**.**

   The Swiss-Prot/TrEMBL Retrieve a List of Entries page appears.

2. **Leave the File Name field blank so the browser returns the sequences directly.**

3. **On the Format line, select the FASTA check box.**

4. **Enter the accession numbers of your sequences in the Sequence window, as shown in Figure 9-5.**

   Enter one accession number (or sequence name) per line. For our example, we entered P20472, P80079, P02626, P02619, P43305, P32930, Q91482, P02620, P02622.

   TIP

   The Upload a File field is convenient if you have all your accession numbers in a file. You can generate this file with Microsoft Word, but it must be in text form. You can enter the filename in the box or use the Browse button and choose your file for upload.

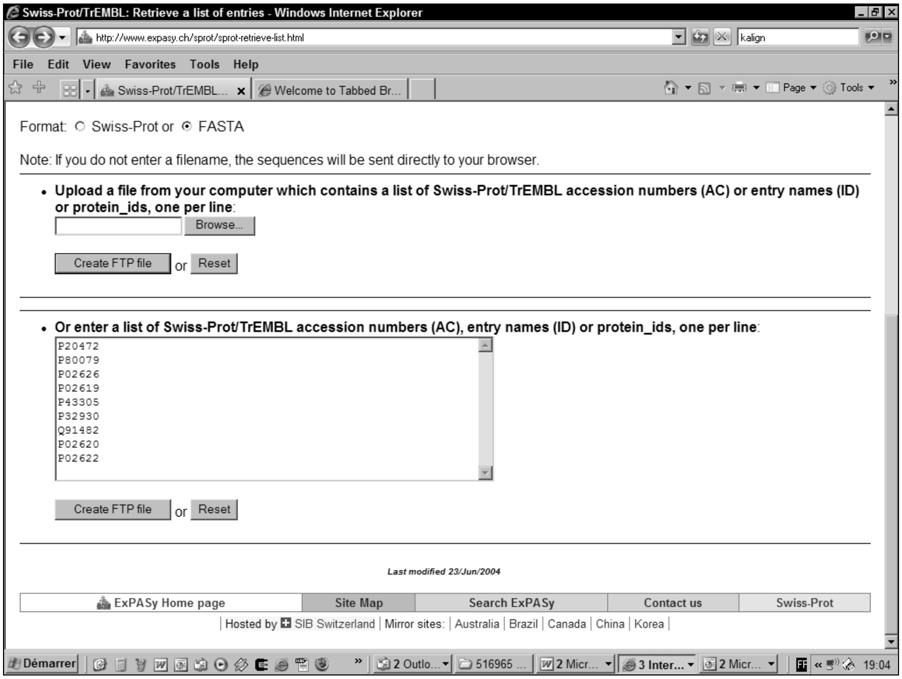
5. **Underneath the field where you entered your accession numbers, click the Create FTP File button.**

   This submits your sequence request to the ExPASy server.

6. **Save the results into a text file with the File➪Save As option of your browser.**

   WARNING!

   This server gives the sequences names that are longer than 15 characters. These can give you trouble on some servers.



Format: ○ Swiss-Prot or ● FASTA

Note: If you do not enter a filename, the sequences will be sent directly to your browser.

- **Upload a file from your computer which contains a list of Swiss-Prot/TrEMBL accession numbers (AC) or entry names (ID) or protein_ids, one per line**:

- **Or enter a list of Swiss-Prot/TrEMBL accession numbers (AC), entry names (ID) or protein_ids, one per line**:

```
P20472
P80079
P02626
P02619
P43305
P32930
Q91482
P02620
P02622
```

Last modified 23/Jun/2004

Hosted by SIB Switzerland | Mirror sites: | Australia | Brazil | Canada | China | Korea |

Figure 9-5: Gathering several sequences from the ExPASy server.

Choosing the Right Method of Multiple Sequence Alignment

Before you start making multiple sequence alignments, you must know that none of the methods available today is perfect. They all use approximations.

Building a multiple alignment that lets you make a real discovery requires some practice. The usual strategy requires comparing several alternative results and looking for robustness and stability.

In this section, we show you how to use ClustalW, the most commonly used multiple sequence alignment package. It is a simple, no-fuss, no-questions-asked kind of tool. We also show you how to use Tcoffee, one of the latest multiple-sequence-alignment packages that you can use. With Tcoffee, you can combine sequences and structures, evaluate an alignment, or merge several alternative multiple alignments into a single unified result. Finally, we also introduce you to MUSCLE, one of the fastest alignment methods around.

Using ClustalW

ClustalW is by far the most commonly used program for making multiple sequence alignments. If you see a multiple alignment in a scientific publication, you can safely bet that the authors used ClustalW to generate it.

ClustalW uses a progressive method to build its alignments. Instead of aligning all the sequences at the same time, it adds them one by one. If you want to get best results from ClustalW, understanding its underlying principle helps a lot.

Many ClustalW servers are around. They usually run the same version of this program, but their interfaces give you access to different options. At the end of this chapter, we give you a list of servers that run ClustalW. Shopping around to find a ClustalW server that's both fast *and* reliable is worth your time.

Running the EBI ClustalW server

ClustalW is a typical example of a program that can produce a reasonable output with its default settings. You should probably worry about the settings only when you want to change the output format of the program.

ClustalW behind the scenes

ClustalW is the latest of the Clustal software series. Clustal was the first multiple sequence alignment program that could run on almost any platform. Legend has it that Des Higgins made the original design for Clustal on the back of an envelope, in a smoky Dublin pub, in the early 1990s. A few years later, when Des moved to the European Molecular Biology Laboratory in Heidelberg, he took Clustal along — and kept it alive, turning it into one of the most successful series of bioinformatics programs: ClustalV, ClustalW, and ClustalX, an X-window implementation.

Des Higgins did not truly invent the Clustal algorithm; Paula Hogeweg had already described it in the early 1980s. What Des did was to put together all the right ingredients so his program could be used by anybody on almost any computer. He also packaged and arranged everything so it would be very easy to use. These are some of the reasons why Clustal became an instant hit. These days, with more than 35,000 citations, ClustalW is one of the most widely cited scientific publications in the history of biology.

ClustalW uses a progressive algorithm. This means that it builds the alignment progressively, adding sequences little by little until the complete multiple sequence alignment is finished. The reason for doing this is that state-of-the art sequence-alignment programs find it hard to align more than two sequences at a time. Since these pairwise-alignment programs are the only tools we've got on hand to produce a multiple alignment, the only solution is to cheat a little. This is what the progressive algorithm does.

The trick is simple: You start comparing all your sequences two by two, so you can cluster them by similarity (Clustal does this for you). The clustering looks like a phylogenetic tree. (It is the file with a `.dnd` extension that ClustalW outputs.)

This clustering is named a *dendrogram.* If we were to align four sequences A, B, C, and D, the dendrogram might look like this:

```
        |------A
 |------|
 |      |------B
 |
Root ----------|
 |
 |      |------C
 |------|
        |------D
```

The topology of this dendrogram tells us a simple story: It says that A and B are more similar to each other than they are to C and D. Thus if we align A with B, we are less likely to make a mistake than if we align A with C or D.

To make the progressive alignment, ClustalW follows the dendrogram topology: It starts aligning A with B. After this it aligns C with D. When this is done, Clustal has two small multiple alignments (AB and CD). This is where Clustal pulls out its main trick: It aligns the two alignments as if each of them was a single sequence! It is not as complicated as it seems and there are many ways to do this. For instance, you could replace each alignment with a single consensus sequence. Clustal uses a slightly more sophisticated method, but the idea is essentially the same: It treats multiple alignments like single sequences and aligns them two by two.

Now you may ask, where do we cheat then? The answer is simple: we make a multiple alignment with ALL the sequences in our set, but we do not use ALL the information they contain. For instance, when Clustal aligns A and B, it does not use C and D. This could be a problem. If A and B are very different, we will produce an incorrect pairwise alignment — which would be a waste if the two other sequences contained some useful information we did not use. Imagine, for instance, that A, B, C, and D all contain a certain very short (but important) motif. This motif does not look so important when you compare only A and B, and it only shows up when you look at all the sequences simultaneously. Unfortunately, with a progressive strategy, C and D will come too late to rescue the incorrect alignment of A and B.

The reason not to use all the information is that it's too expensive in terms of computation. So we cheat a little and use the progressive alignment. This shouldn't worry you too much, though. Even if it is a little greedy and approximate, Clustal often delivers pretty good alignments.

When you ask yourself which type of sequences are best suited for ClustalW, imagine that your sequences are like big stones spread across a shallow river. Making a multiple alignment is like crossing this river by jumping from stone to stone: It doesn't matter how wide the river is, as long as you always find a stone to jump to. Similarly, it doesn't matter how many sequences you have, and how far apart they are from each other, as long as a chain of correct alignments exists that can take you across the entire set. (Of course, if a sequence or a small group of sequences is very different from the rest, you fall in the rapids!)

Sometimes your sequence set may contain many identical or similar sequences — usually a problem because sequences that belong to minority subgroups become harder to align properly. If you can, you want to avoid this situation by removing the sequences yourself, but if you have no choice, it can be reassuring to know that ClustalW is equipped to deal with redundancy.

In case you were wondering, the *W* in *ClustalW* does not stand for Dubya, the U.S. President; it stands for *W*eights. ClustalW uses a sophisticated scheme so that very similar sequences do not end up dominating the multiple sequence alignment. In fact, every sequence is supposed to receive a weight proportional to the amount of new information it contributes.

Before you head off to a ClustalW server, you must do a little spadework ahead of time. Specifically, you need to gather together all the sequences you want to work with. (If you're not quite sure how to do this, check out the section "Gathering a known collection of sequences from Swiss-Prot," earlier in this chapter.)

The most convenient way to use ClustalW is to feed it sequences in FASTA format. However, you can also give to ClustalW a variety of formats, including Swiss-Prot and PIR (the Protein Information Resource format), as well as sequences in the most common multiple-alignment formats.

If you want to use ClustalW effectively, you'll need to observe a couple of caveats:

- **If you give ClustalW a set of sequences already aligned, it does not remove the existing gaps.** This means that the alignment you input influences the alignment ClustalW will produce.
- **The order in which you give sequences to ClustalW sometimes influences the alignment.** If you change this order, even with the same sequences, the alignment may change.

With your sequences in hand, it's now time to fire up the ClustalW server:

1. **Point your browser to the EBI ClustalW server page at** `www.ebi.ac.uk/clustalw`**.**

   The ClustalW page dutifully appears.

2. **Paste the sequences you collected in the Sequence window.**
3. **Choose Fast from the Alignment pull-down menu (Figure 9-6).**
4. **Use the Output Format pull-down menu to set the selection of your choice.**

   Output formats have various pros and cons. (See Chapter 10 for a discussion on this.) It is safe to use Aln Without Numbers, the default ClustalW format.

   It is never too late to change a format. If you didn't generate your multiple alignment in the format that suits you best, DON'T recompute it! You can easily reformat alignments by using an online reformat utility (such as Fmtseq) at `www.bimcore.emory.edu/Pise/`. (For more on reformatting, see Chapter 10.)

5. **Choose Input from the Output Order pull-down menu. (Refer to Figure 9-6.)**

   Here are a few things to remember about setting output:

- When you set the output order to Input, ClustalW outputs your sequences in their original order.
- If you set the output order to Aligned, the sequences appear in the order they were aligned. This order depends on the guide tree, and it puts closely related sequences next to each other.
- Although the Aligned output is more informative than Input, Aligned output makes it difficult to compare alignments generated with different methods or different parameters. Usually it's better to prearrange your sequences in the most informative way and *then* select the Input option.

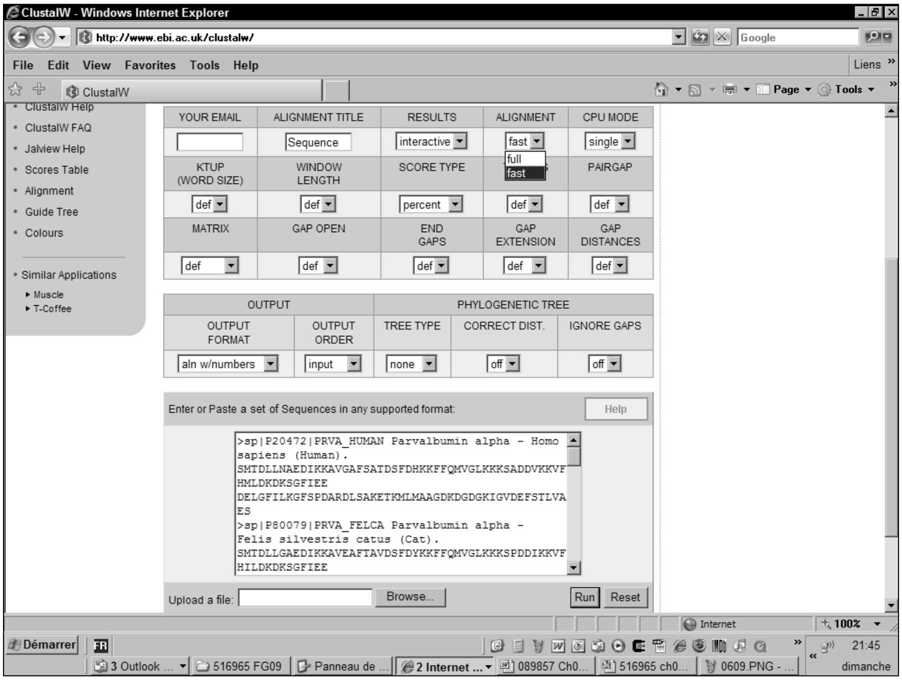
WARNING!

6. **Do not select a tree type.**

   No need to select a tree type here. Any tree computed would be estimated from the unaligned sequences. If you want to compute a tree, compute the alignment first and then turn that alignment into a phylogenetic tree. (See Chapter 13 for more on phylogenetic trees.)

7. **Click the Run button at the bottom of the page.**

   An intermediate page appears. Wait until your browser displays the Results page.



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
	Sequence	interactive	fast	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	full fast	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	input	none	off	off

Enter or Paste a set of Sequences in any supported format:

```
>sp|P20472|PRVA_HUMAN Parvalbumin alpha - Homo
sapiens (Human).
SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVF
HMLDKDKSGFIEE
DELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVA
ES
>sp|P80079|PRVA_FELCA Parvalbumin alpha -
Felis silvestris catus (Cat).
SMTDLLGAEDIKKAVEAFTAVDSFDYKKFFQMVGLKKKSPDDIKKVF
HILDKDKSGFIEE
```

Upload a file:

Figure 9-6: The EBI ClustalW server.

8. **Save your results.**

   Your results come in three sections:

   - **Pairwise scores:** This is the default output of ClustalW. You can safely ignore the content of this section. It has mostly to do with the pairwise comparisons that ClustalW makes to figure out the guide tree.
   - **The multiple alignment:** This section is in the middle of the output. It contains your alignment. You can display this alignment and save it as a text file if you click the hyperlink that comes just before the alignment.

     Most ClustalW Web servers (including the EBI) output the ClustalW guide tree (a `.dnd` file). The guide tree is NOT a phylogenetic tree. To obtain a phylogenetic tree from a ClustalW server, you must cut and paste an actual multiple sequence alignment — not a set of unaligned sequences. (We explain how to do this in Chapter 13.)

Changing ClustalW parameters

There are three parameters in ClustalW that can change your alignment: substitution matrices, gap-opening penalties, and gap-extension penalties. (Refer to Figure 9-6 to see the pull-down menus associated with these parameters.) Table 9-5 lists these three types of parameters along with potential consequences when you change them.

Table 9-5 Controlling the Effect of Parameter Tuning in ClustalW

Parameter	*Effect*
Substitution matrix	Substitution matrices control the cost of mutations in sequence alignments. (For more on cost, see Chapter 8.) If you select a category of matrices like PAM or BLOSUM, ClustalW automatically chooses the most adapted index. Predicting the effect of changing matrices is difficult, and there is no such thing as an ideal matrix. If your sequences are closely related, such a change has no effect. If your alignment is difficult to interpret, it may be worth changing from BLOSUM to PAM.
Gap-opening penalty	Gap-opening penalties (GOP) control the cost of opening gaps in your alignment. The higher the value, the more difficult it is to insert a gap in your alignment. The gap-opening penalty is applied once for the opening of each gap. Tuning has little effect because ClustalW readjusts these values automatically.
Gap-extension penalty	Gap-extension penalties control the size of the gaps. It's impossible to predict the optimal couple GOP/GEP, but it's clear that an optimal value exists for almost every protein family — and that the only way to find this value is empirically.

WARNING!

Never change parameters in order to force ClustalW to produce an alignment that you know is right. If you know the correct alignment, edit the ClustalW alignment with one of the editors we show you in Chapter 10. The only good justification for changing the parameters is to see whether slight changes can improve the overall appearance of your alignment in terms of blocks and weakly conserved positions.

Aligning sequences and structures with Tcoffee

Tcoffee is a recent method developed for conducting multiple sequence alignments. It uses a principle that's a bit similar to ClustalW, but it yields more accurate alignments at the cost of a slightly longer running time. Tcoffee builds a progressive alignment like ClustalW, but it compares segments across the entire sequence set. Table 9-6 lists the main applications of Tcoffee. Aside from its accuracy, the main specificity of Tcoffee is its ability to align sequences and structures (EXPRESSO), the possibility of evaluating the accuracy of an alignment (CORE) and the possibility of combining many alternative multiple sequence alignments into one (Mcoffee).

Table 9-6 Available Tools on www.tcoffee.org

Usage	*Description*
TCOFFEE	Produce a multiple sequence alignment with Tcoffee.
CORE	Evaluate the reliability of an existing multiple alignment.
MCOFFEE	Run any requested Multiple Sequence Alignment package and combine all the output into one final alignment.
EXPRESSO	Incorporate all the available structural information in your alignment. Will produce the best sequence alignments if the structures are available.

Making a multiple alignment with Tcoffee

Making a Multiple sequence alignment with the regular Tcoffee server is only a matter of cutting and pasting your sequences into the right window. (If you're not quite sure how to get the sequences, check out the section "Gathering a known collection of sequences from Swiss-Prot," earlier in this chapter.)

With your sequences in hand, it's Coffee Time!

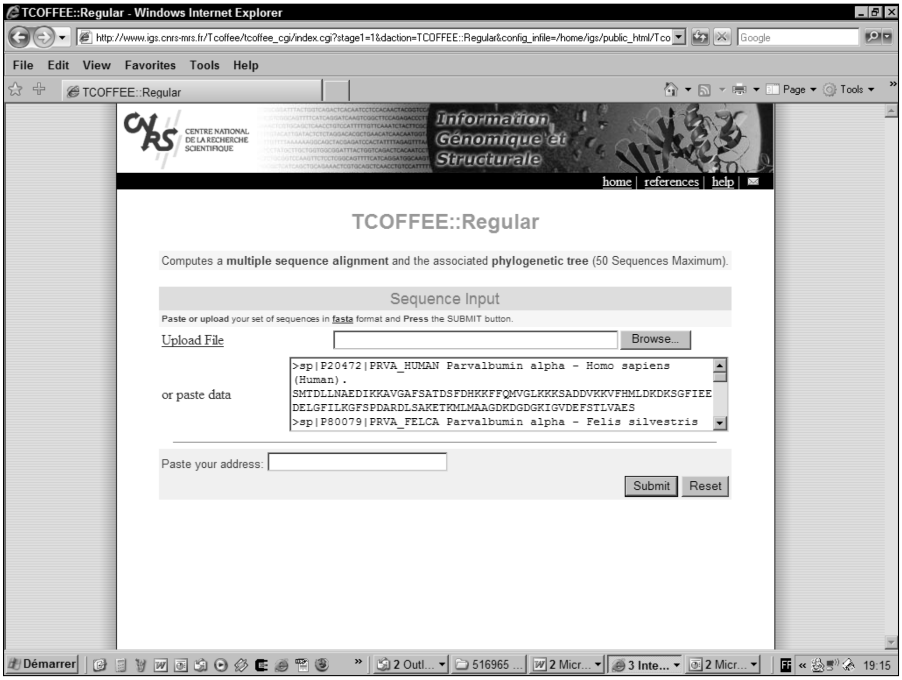
1. **Point your browser to the Tcoffee server home page at** `www.tcoffee.org`.
2. **Click the Regular button on the TCOFFEE line (first line).**

   The Build a Multiple Alignment page appears (Figure 9-7).
3. **Paste your sequences into the large window.**

   You can use most formats. If your sequences are in a text file, you can upload this file by using the Browse button.
4. **Click the Submit button at the top or the bottom of the page.**

   Tcoffee can be slow at times. If you'd prefer to be notified when your computation is done, enter your e-mail address in the Web form.
5. **Examine your results.**

   Tcoffee returns a table that contains hyperlinks to your results, as shown in Figure 9-8.


TCOFFEE::Regular - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee/tcoffee_cgi/index.cgi?stage1=1&daction=TCOFFEE::Regular&config_infile=/home/igs/public_html/Tco
Google
File
Edit
View
Favorites
Tools
Help
TCOFFEE::Regular
Page
Tools
CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
Information Génomique et Structurale
home
references
help
Démarrer
2 Outl...
516965 ...
2 Micr...
3 Inte...
2 Micr...
19:15

TCOFFEE::Regular

Computes a **multiple sequence alignment** and the associated **phylogenetic tree** (50 Sequences Maximum).

Sequence Input

Paste or upload your set of sequences in fasta format and **Press** the SUBMIT button.

Upload File [ ] Browse...

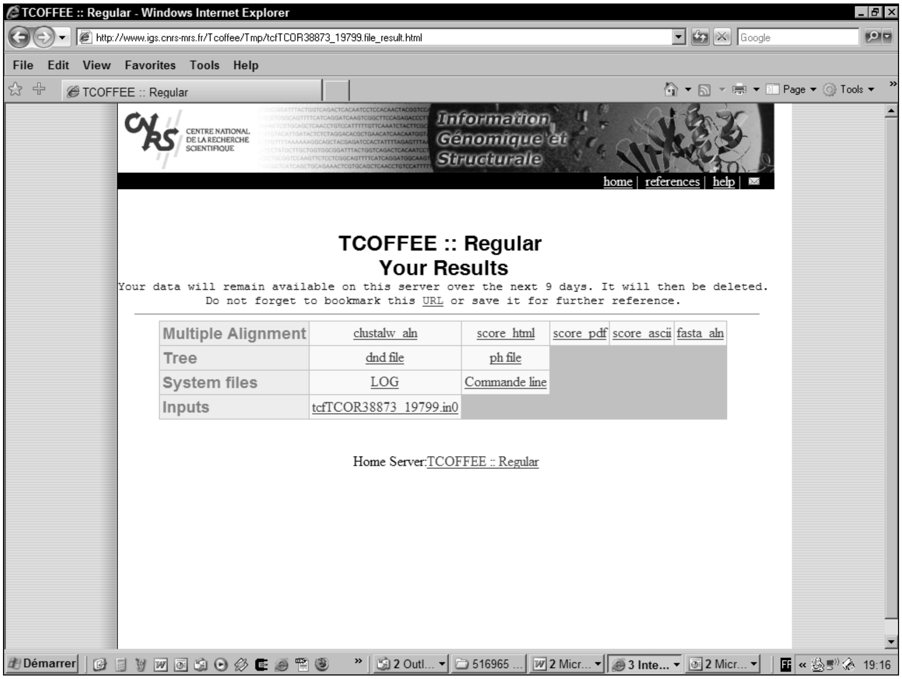
or paste data

```
>sp|P20472|PRVA_HUMAN Parvalbumin alpha - Homo sapiens
(Human).
SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDKSGFIEE
DELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES
>sp|P80079|PRVA_FELCA Parvalbumin alpha - Felis silvestris
```

Paste your address: [ ]

Submit Reset

Figure 9-7: The Tcoffee Build a Multiple Alignment page.


TCOFFEE :: Regular - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR38873_19799.file_result.html
Google
File Edit View Favorites Tools Help
TCOFFEE :: Regular
Page
Tools
CNRS CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE
Information Génomique et Structurale
home | references | help |
Démarrer
2 Outl...
516965 ...
2 Micr...
3 Inte...
2 Micr...
19:16

TCOFFEE :: Regular
Your Results

Your data will remain available on this server over the next 9 days. It will then be deleted.
Do not forget to bookmark this URL or save it for further reference.

Multiple Alignment	clustalw_aln	score_html	score_pdf	score_ascii	fasta_aln
Tree	dnd file	ph file			
System files	LOG	Commande line			
Inputs	tcfTCOR38873_19799.in0				

Home Server:TCOFFEE :: Regular

Figure 9-8: The Tcoffee default output.

The first row of the table is dedicated to multiple sequence alignments and includes

- **msf_aln, clustalw_aln, fasta_aln:** Text files containing your alignment in various formats.

  Keep these files if you want to use your alignment as input for another program.

- **score_html, score_ascii:** A colorized alignment where every residue appears on a background that indicates the quality of this alignment. Red indicates high-quality segments; blue indicates regions of your alignment that you have no reason to trust. The `score_ascii` is a text version of the `.html` file.

  These two last files are meant only for display; you can't use them as an input for other sequence-analysis programs.

The second row is dedicated to phylogenetic trees:

- **dnd:** The guide tree or dendrogram generated by Tcoffee in Newick format (see Chapter 13). You should not use it in place of the true phylogenetic tree.

- **phylogenetic_tree:** The true phylogenetic tree in Newick format, generated from the Tcoffee multiple alignment by using the Neighbor Joining method (see Chapter 13). This is not a guide tree but a real phylogenetic tree.
- **pdf:** A pdf picture of the phylogenetic tree that corresponds to the `phylogenetic_tree` file.

Combining sequences and structures with EXPRESSO

EXPRESSO is the latest development of Tcoffee, replacing what was known as 3D-Coffee. When you run Expresso, the program uses BLAST to search the PDB (database of structures) for structures whose sequences are similar to your sequences. It then uses theses structures to guide the alignment. Alignments based on structures are expected to be much more accurate than simple sequence alignments.

EXPRESSO is slower than Tcoffee, but if it finds enough structures it produces the most accurate sequence alignments available today, so it's worth a try! EXPRESSO aligns the structures using SAP, a program from Taylor and Orengo, and it aligns sequences and structures using FUGUE, a threading package from Kenji Mizuguchi (developed in Tom Blundell's lab at Cambridge University).

To run EXPRESSO, simply click on the Regular button of the EXPRESSO line on `www.tcoffee.org`. The rest is identical to Tcoffee — so identical that as a user you'd never know you're using structures!

In the output section, look for a file named `template_list`. It lists every structure that the program managed to associate with your original sequences. If this file is empty, it means no structure was available to align your sequences and it means your EXPRESSO alignment was merely a standard Tcoffee alignment.

Evaluating the quality of an alignment with CORE

If you want, you can give Tcoffee a multiple alignment that you generated with your favorite method (or by hand if you are a specialist), and you can ask Tcoffee to evaluate the quality of this multiple sequence alignment for you — which can give you an idea of which portions of your alignment you can trust and which are safer to ignore. To do so, cut and paste your alignment into the CORE server on `www.tcoffee.org`. You can use any of the most common formats (MSF, ALN, FASTA, and PIR).

These evaluations are only empirical, and they do not replace E-values. Nonetheless, it's useful to know that residues with a yellow/orange/red background have an index above 5 (out of 10) — and have more than an 80 percent chance of being correctly aligned.

Tcoffee and sequence democracy

Tcoffee is a progressive alignment method and, in many ways, a proud relative of ClustalW. The main difference between Tcoffee and ClustalW is that Tcoffee doesn't directly use substitution matrices to align sequences; it's much lazier, and simply relies on other methods to work for it.

When you give sequences to Tcoffee, it starts making pairwise comparisons. To do so, it takes every possible pair of sequences and makes a global alignment with ClustalW. It also makes a Lalign comparison for each of these pairs. (Lalign is a local-alignment method developed by Huang and Miller to do pairwise local alignments.) Given two sequences, Lalign produces the ten best local alignments. The complete collection of local and global alignments is a *library.*

After the library is finished, Tcoffee builds a multiple alignment that has the highest possible agreement with all the pairwise alignments in the library. It does so by using a progressive algorithm like ClustalW. This process looks much like a general election, where all the bits of information contained in the library are competing to find their way into the final alignment.

The nice thing about the libraries is that they can contain whatever you want: pairwise alignments, multiple alignments (Mcoffee), global or local, structure-based sequence alignments (EXPRESSO), or even alignments you made yourself by using experimental information. Tcoffee also makes it possible to measure the local agreement between your multiple sequence alignment and any library (CORE).

Crunching large datasets with MUSCLE

MUSCLE is a newcomer in the multiple-sequence-alignment arena — but it is a remarkably efficient package for making fast, high-quality multiple sequence alignments. MUSCLE is ideal if you want to align several hundred sequences. You can access it on various servers, including its home page (at `www.drive5.com/muscle/`). Running MUSCLE is very straightforward — only a matter of cutting and pasting your sequences into the designated window.

Interpreting Your Multiple Sequence Alignment

Interpreting an alignment is a bit of an art. E-values — the scores that tell you how reliable your database search is — do not (yet) exist for multiple sequence alignments. And that means deciding whether your alignment is correct still involves some educated guesswork.

DNA alignments are by far the most difficult to interpret. If you're analyzing this type of sequence, you want a very high level of conservation, knowing that single conserved columns are likely to be meaningless. A DNA block is only informative when it contains several identical columns in a cluster. Even with the DNA of closely related sequences, obtaining such an alignment is still difficult. This is why most biologists (when they have a choice) much prefer protein alignments.

Recognizing the good parts in a protein alignment

The most convincing evaluative grid we have for a protein multiple alignment stems from our knowledge of protein structures (see Chapter 11). We know that structures contain surface loops that evolve rapidly. Loops are softer portions of the protein that connect its more rigid portions. Protein structures also contain core regions that act as support walls for the protein. These support walls evolve less rapidly than the loops on the surface (see Chapter 11).

In your multiple alignment, you can expect to find nice, gap-free blocks that correspond to the core regions — and gap-rich regions that correspond to the loops. The alignment shown in Figure 9-9 (which appears in the following section, "Taking your multiple alignment further") is a good illustration of this principle.

Now, how can you tell whether a block is good? When you look at a ClustalW, a MUSCLE, or a Tcoffee alignment, you can see that the last line contains seemingly-cabalistic signs such as (*), (:), or (.). These three symbols have very precise meanings:

- (*) A star indicates an entirely conserved column.
- (:) A colon indicates columns where all the residues have roughly the same size and the same hydropathy.
- (.) A period indicates columns where the size *or* the hydropathy has been preserved in the course of evolution.

After you get used to them, these indications are really quite priceless. For instance, your average *good block* is a unit at least 10–30 amino acids long, exhibiting at least one to three stars (*), a few more colons (:) close to the stars, and a several periods (.) sprinkled here and there. This is exactly what you see in the alignment shown in Figure 9-9.

The magic thing about multiple sequence alignments is that 4 or 5 conserved positions over 50 amino acids can be enough to convince us that we're looking at a genuine signal. This is less than 10 percent identity! If you remember that we require at least 25 percent identity to consider a pairwise alignment interesting, you can easily understand why multiple alignments are so powerful. They can help you find treasure in the twilight zone, that region of similarity where pair-wise identity does not tell you anything about potential homology (see Chapter 8).

Another criterion for a useful multiple alignment is knowing the type of amino acids you can expect to see conserved. Amino acids aren't equal — and they all have very characteristic patterns of mutation/conservation in a multiple sequence alignment. Table 9-7 lists the most common features associated with some conserved columns you may come across.

Conserved columns in a multiple sequence alignment are meaningful only when the surrounding columns are not conserved.

Table 9-7 Patterns of Conservation in Multiple Sequence Alignments

Amino Acid	*Characteristic*
W,Y, F	It is common to find conserved tryptophans. Tryptophan is a large hydrophobic residue that sits deep in the core of proteins. It plays an important role in their stability and is therefore difficult to mutate. When tryptophan mutates, it is usually replaced by another aromatic amino acid, such as phenylalanine or tyrosine. Patterns of conserved aromatic amino acids constitute the most common signatures for recognizing protein domains.
G, P	It is common to find conserved columns with a glycine or a proline in a multiple alignment. These two amino acids often coincide with the extremities of well-structured beta strands or alpha helices. (For more on these structures, see Chapter 11.)
C	Cysteines are famous for making C-C (disulphide) bridges. Conserved columns of cysteines are rather common and usually indicate such bridges. Columns of conserved cysteines with a specific distance provide a useful signature for recognizing protein domains and folds.
H, S	Histidine and serine are often involved in catalytic sites, especially those of proteases. Conserved histidine or a conserved serine are good candidates for being part of an active site.
K, R, D, E	These charged amino acids are often involved in ligand binding. Highly conserved columns can also indicate a salt bridge inside the core of the protein.
L	Leucines are rarely very conserved unless they're involved in protein-protein interactions such as a leucine zipper.

Taking your multiple alignment further

One sentence summarizes what you really want from your multiple alignment: You want to identify important positions! You want to find the amino acids that are not allowed to mutate, those that you find conserved even when aligning distantly related proteins.

Consider, for instance, the alignment in Figure 9-9. You can use any ClustalW or Tcoffee server to generate this multiple sequence alignment using any server you fancy. It is a good alignment: It contains distantly related proteins, and it is beginning to tell us a nice story about the various components of our protein family. For instance, we can clearly see that the N-terminus region seems to be more conserved than the C-terminus region. In the N-terminus, we can see a short stretch of highly conserved amino acids that make this region a good candidate for being a binding or an active site.

This is interesting — but it isn't enough to make a big story. This alignment still contains too many conserved positions for a detailed analysis. At this point, what we could do is add a few distantly related sequences, one by one, and carefully check the effect of these sequences on the overall alignment quality. More specifically, we want to make sure that these distantly related sequences actually enhance existing patterns rather than completely destroying blocks.

Here is a possible strategy to further reveal the evolutionary constraints within our protein. This strategy relies on the integration within the multiple alignment of precisely those sequences that BLAST reported as marginal hits when we first scanned Swiss-Prot for homologues of the human parvalbumin. (If this all sounds a bit unfamiliar, take a look at the "Selecting sequences on the ExPASy server" section, earlier in this chapter.) You don't actually need to rerun BLAST to use this example; we give you the info you need to know. First and foremost, gather your sequences as follows:

1. **Point your browser to** `www.expasy.ch/sprot/sprot-retrieve-list.html`**.**

   The Swiss-Prot/TrEMBL: Retrieve a List of Entries page appears.

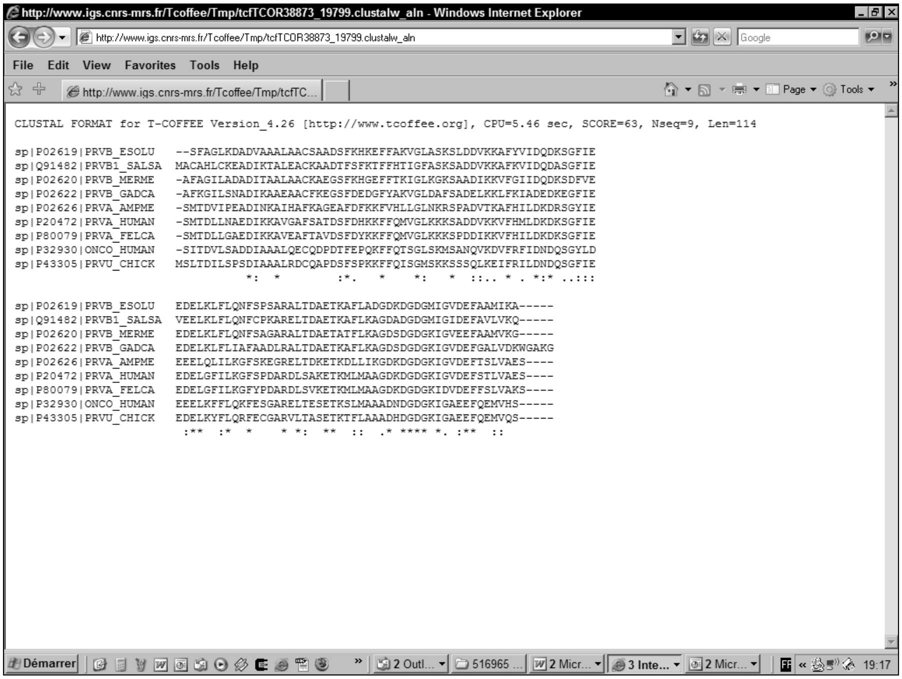
2. **In the Format line, select the FASTA radio button.**

3. **Enter the accession number of your sequences in the Sequence window.**

   Enter one accession number per line. For our example, we entered P20472, P80079, P02626, P02619, P43305, P32930, Q91482, P02620, P02622, P02586.

   P02586 is the TPCS_RABIT, the troponin C of rabbit. In the BLAST of the human parvalbumin against Swiss-Prot that we used to select the other sequences, BLAST reported this hit with an E-value of 5!

   On its own, this result is not interesting — but now that we have a multiple sequence alignment, we can see whether this rabbit can tell us something about our human protein.


http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR38873_19799.clustalw_aln - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR38873_19799.clustalw_aln
Google
File Edit View Favorites Tools Help
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTC...
Page
Tools
Démarrer
2 Outl...
516965 ...
2 Micr...
3 Inte...
2 Micr...
19:17

```
CLUSTAL FORMAT for T-COFFEE Version_4.26 [http://www.tcoffee.org], CPU=5.46 sec, SCORE=63, Nseq=9, Len=114

sp|P02619|PRVB_ESOLU   --SFAGLKDADVAAALAACSAADSFKHKEFFAKVGLASKSLDDVKKAFYVIDQDKSGFIE
sp|Q91482|PRVB1_SALSA  MACAHLCKEADIKTALEACKAADTFSFKTFFHTIGFASKSADDVKKAFKVIDQDASGFIE
sp|P02620|PRVB_MERME   -AFAGILADADITAALAACKAEGSFKHGEFFTKIGLKGKSAADIKKVFGIIDQDKSDFVE
sp|P02622|PRVB_GADCA   -AFKGILSNADIKAAEAACFKEGSFDEDGFYAKVGLDAFSADELKKLFKIADEDKEGFIE
sp|P02626|PRVA_AMPME   -SMTDVIPEADINKAIHAFKAGEAFDFKKFVHLLGLNKRSPADVTKAFHILDKDRSGYIE
sp|P20472|PRVA_HUMAN   -SMTDLLNAEDIKKAVGAFSATDSFDHKKFFQMVGLKKKSADDVKKVFHMLDKDKSGFIE
sp|P80079|PRVA_FELCA   -SMTDLLGAEDIKKAVEAFTAVDSFDYKKFFQMVGLKKKSPDDIKKVFHILDKDKSGFIE
sp|P32930|ONCO_HUMAN   -SITDVLSADDIAAALQECQDPDTFEPQKFFQTSGLSKMSANQVKDVFRFIDNDQSGYLD
sp|P43305|PRVU_CHICK   MSLTDILSPSDIAAALRDCQAPDSFSPKKFFQISGMSKKSSSQLKEIFRILDNDQSGFIE
                                 *:  *        :*.   *    *:   *  ::.. * . *:* ..:::

sp|P02619|PRVB_ESOLU   EDELKLFLQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-----
sp|Q91482|PRVB1_SALSA  VEELKLFLQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-----
sp|P02620|PRVB_MERME   EDELKLFLQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-----
sp|P02622|PRVB_GADCA   EDELKLFLIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG
sp|P02626|PRVA_AMPME   EEELQLILKGFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAES----
sp|P20472|PRVA_HUMAN   EDELGFILKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----
sp|P80079|PRVA_FELCA   EDELGFILKGFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----
sp|P32930|ONCO_HUMAN   EEELKFFLQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----
sp|P43305|PRVU_CHICK   EDELKYFLQRFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----
                        :**  :*  *    * *:  **  ::  .* **** *. :**  ::
```

Figure 9-9: A good multiple sequence alignment.

4. **Click the Create FTP File button.**
5. **Copy the sequences onto the Clipboard.**

When you've got all the sequences in your basket (well, on the Clipboard at least), get ready to do some alignment work:

1. **Point your browser to the EBI ClustalW server home page at** `www.ebi.ac.uk/clustalw/index.html`.
2. **Paste the sequences you gathered in the preceding steps list into the Sequence window.**
3. **Choose Fast from the Alignment pull-down menu.**
4. **Use the Output Format pull-down menu to choose the output format you want.**
5. **Choose Input from the Output Order pull-down menu.**
6. **Click the Run button at the bottom of the page.**
7. **Save your results.**

You can see in Figure 9-10 that the new sequence respects the blocks that already existed (Figure 9-9), while shunting some conserved positions. It also reveals regions where insertion and deletions are likely to occur. These regions are good candidates for being loops. (Remember that loops are portions of the protein structure that are a bit more flexible than the rest of the structure and act as connectors; see Chapter 11 for more details.)

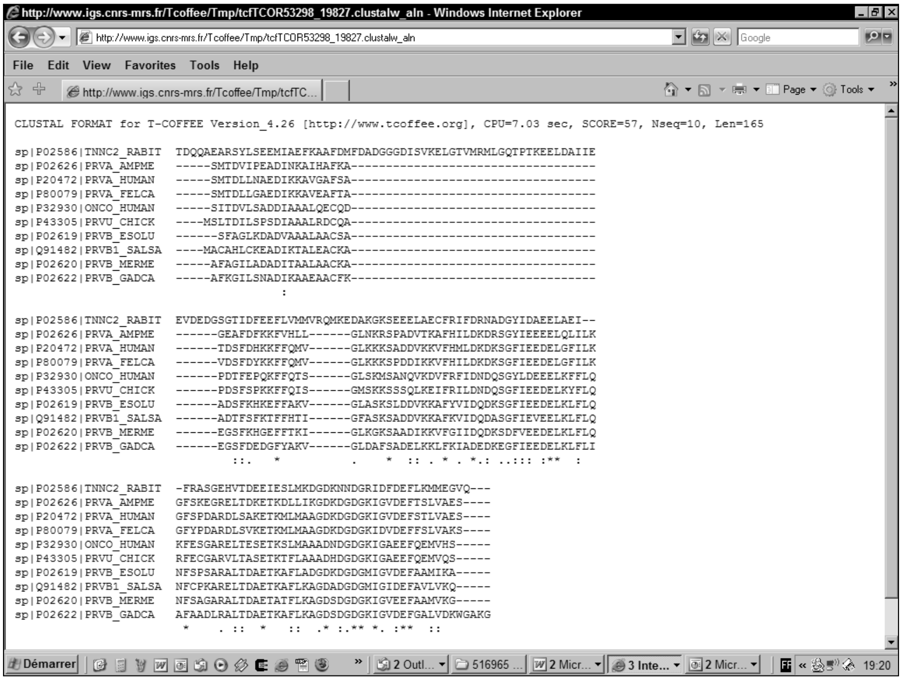
Figure 9-10 also reveals the existence of two conserved regions in the N-terminus of the human parvalbumin. This is an interesting result that narrows down the range of possibilities if we were looking for the amino acids responsible for the function of our protein.

At this point, going a bit further and adding another distantly related protein is a good idea. Our aim is to check that these few highly conserved regions are indeed conserved across the whole protein family, even when we compare distantly related cousins.

To do this check, you can go once again through the two previous step lists, using an extra sequence:

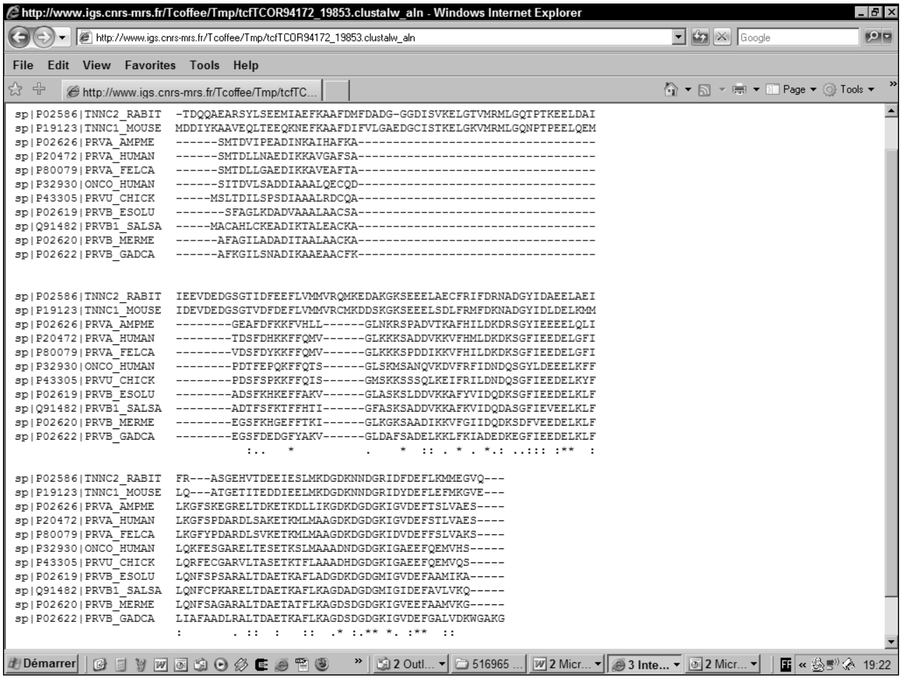
> P20472, P80079, P02626, P02619, P43305, P32930, Q91482, P02620, P02622, P02586, P19123

P19123 is TPCC_MOUSE, the mouse troponin C. It is a very remote homologue of the human parvalbumin. Figure 9-11 shows the result of the inclusion of this new protein in our multiple sequence alignment. It clearly shows that most conserved columns remain.


http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR53298_19827.clustalw_aln - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR53298_19827.clustalw_aln
Google
File
Edit
View
Favorites
Tools
Help
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTC...
Page
Tools
Démarrer
2 Outl...
516965 ...
2 Micr...
3 Inte...
2 Micr...
FR
19:20

```
CLUSTAL FORMAT for T-COFFEE Version_4.26 [http://www.tcoffee.org], CPU=7.03 sec, SCORE=57, Nseq=10, Len=165

sp|P02586|TNNC2_RABIT  TDQQAEARSYLSEEMIAEFKAAFDMFDADGGGDISVKELGTVMRMLGQTPTKEELDAIIE
sp|P02626|PRVA_AMPME   -----SMTDVIPEADINKAIHAFKA-----------------------------------
sp|P20472|PRVA_HUMAN   -----SMTDLLNAEDIKKAVGAFSA-----------------------------------
sp|P80079|PRVA_FELCA   -----SMTDLLGAEDIKKAVEAFTA-----------------------------------
sp|P32930|ONCO_HUMAN   -----SITDVLSADDIAAALQECQD-----------------------------------
sp|P43305|PRVU_CHICK   ----MSLTDILSPSDIAAALRDCQA-----------------------------------
sp|P02619|PRVB_ESOLU   ------SFAGLKDADVAAALAACSA-----------------------------------
sp|Q91482|PRVB1_SALSA  ----MACAHLCKEADIKTALEACKA-----------------------------------
sp|P02620|PRVB_MERME   -----AFAGILADADITAALAACKA-----------------------------------
sp|P02622|PRVB_GADCA   -----AFKGILSNADIKAAEAACFK-----------------------------------
                                      :

sp|P02586|TNNC2_RABIT  EVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDAEELAEI--
sp|P02626|PRVA_AMPME   ------GEAFDFKKFVHLL------GLNKRSPADVTKAFHILDKDRSGYIEEEELQLILK
sp|P20472|PRVA_HUMAN   ------TDSFDHKKFFQMV------GLKKKSADDVKKVFHMLDKDKSGFIEEDELGFILK
sp|P80079|PRVA_FELCA   ------VDSFDYKKFFQMV------GLKKKSPDDIKKVFHILDKDKSGFIEEDELGFILK
sp|P32930|ONCO_HUMAN   ------PDTFEPQKFFQTS------GLSKMSANQVKDVFRFIDNDQSGYLDEEELKFFLQ
sp|P43305|PRVU_CHICK   ------PDSFSPKKFFQIS------GMSKKSSSQLKEIFRILDNDQSGFIEEDELKYFLQ
sp|P02619|PRVB_ESOLU   ------ADSFKHKEFFAKV------GLASKSLDDVKKAFYVIDQDKSGFIEEDELKLFLQ
sp|Q91482|PRVB1_SALSA  ------ADTFSFKTFFHTI------GFASKSADDVKKAFKVIDQDASGFIEVEELKLFLQ
sp|P02620|PRVB_MERME   ------EGSFKHGEFFTKI------GLKGKSAADIKKVFGIIDQDKSDFVEEDELKLFLQ
sp|P02622|PRVB_GADCA   ------EGSFDEDGFYAKV------GLDAFSADELKKLFKIADEDKEGFIEEDELKLFLI
                               ::.   *          .    *  :: . * . *.: ..::: :**  :

sp|P02586|TNNC2_RABIT  -FRASGEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ---
sp|P02626|PRVA_AMPME   GFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAES----
sp|P20472|PRVA_HUMAN   GFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----
sp|P80079|PRVA_FELCA   GFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----
sp|P32930|ONCO_HUMAN   KFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----
sp|P43305|PRVU_CHICK   RFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----
sp|P02619|PRVB_ESOLU   NFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-----
sp|Q91482|PRVB1_SALSA  NFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-----
sp|P02620|PRVB_MERME   NFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-----
sp|P02622|PRVB_GADCA   AFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG
                        *    . ::  *   ::  .* :.** *. :**  ::
```

Figure 9-10: Making a multiple alignment of distantly related proteins.


http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR94172_19853.clustalw_aln - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTCOR94172_19853.clustalw_aln
Google
File Edit View Favorites Tools Help
http://www.igs.cnrs-mrs.fr/Tcoffee/Tmp/tcfTC...
Page
Tools
Démarrer
2 Outl...
516965 ...
2 Micr...
3 Inte...
2 Micr...
FR
19:22

```
sp|P02586|TNNC2_RABIT  -TDQQAEARSYLSEEMIAEFKAAFDMFDADG-GGDISVKELGTVMRMLGQTPTKEELDAI
sp|P19123|TNNC1_MOUSE  MDDIYKAAVEQLTEEQKNEFKAAFDIFVLGAEDGCISTKELGKVMRMLGQNPTPEELQEM
sp|P02626|PRVA_AMPME   ------SMTDVIPEADINKAIHAFKA----------------------------------
sp|P20472|PRVA_HUMAN   ------SMTDLLNAEDIKKAVGAFSA----------------------------------
sp|P80079|PRVA_FELCA   ------SMTDLLGAEDIKKAVEAFTA----------------------------------
sp|P32930|ONCO_HUMAN   ------SITDVLSADDIAAALQECQD----------------------------------
sp|P43305|PRVU_CHICK   -----MSLTDILSPSDIAAALRDCQA----------------------------------
sp|P02619|PRVB_ESOLU   -------SFAGLKDADVAAALAACSA----------------------------------
sp|Q91482|PRVB1_SALSA  -----MACAHLCKEADIKTALEACKA----------------------------------
sp|P02620|PRVB_MERME   ------AFAGILADADITAALAACKA----------------------------------
sp|P02622|PRVB_GADCA   ------AFKGILSNADIKAAEAACFK----------------------------------


sp|P02586|TNNC2_RABIT  IEEVDEDGSGTIDFEEFLVMMVRQMKEDAKGKSEEELAECFRIFDRNADGYIDAEELAEI
sp|P19123|TNNC1_MOUSE  IDEVDEDGSGTVDFDEFLVMMVRCMKDDSKGKSEEELSDLFRMFDKNADGYIDLDELKMM
sp|P02626|PRVA_AMPME   --------GEAFDFKKFVHLL------GLNKRSPADVTKAFHILDKDRSGYIEEEELQLI
sp|P20472|PRVA_HUMAN   --------TDSFDHKKFFQMV------GLKKKSADDVKKVFHMLDKDKSGFIEEDELGFI
sp|P80079|PRVA_FELCA   --------VDSFDYKKFFQMV------GLKKKSPDDIKKVFHILDKDKSGFIEEDELGFI
sp|P32930|ONCO_HUMAN   --------PDTFEPQKFFQTS------GLSKMSANQVKDVFRFIDNDQSGYLDEEELKFF
sp|P43305|PRVU_CHICK   --------PDSFSPKKFFQIS------GMSKKSSSQLKEIFRILDNDQSGFIEEDELKYF
sp|P02619|PRVB_ESOLU   --------ADSFKHKEFFAKV------GLASKSLDDVKKAFYVIDQDKSGFIEEDELKLF
sp|Q91482|PRVB1_SALSA  --------ADTFSFKTFFHTI------GFASKSADDVKKAFKVIDQDASGFIEVEELKLF
sp|P02620|PRVB_MERME   --------EGSFKHGEFFTKI------GLKGKSAADIKKVFGIIDQDKSDFVEEDELKLF
sp|P02622|PRVB_GADCA   --------EGSFDEDGFYAKV------GLDAFSADELKKLFKIADEDKEGFIEEDELKLF
                                 :..   *          .    *  :: . * . *.: ..::: :**  :

sp|P02586|TNNC2_RABIT  FR---ASGEHVTDEEIESLMKDGDKNNDGRIDFDEFLKMMEGVQ---
sp|P19123|TNNC1_MOUSE  LQ---ATGETITEDDIEELMKDGDKNNDGRIDYDEFLEFMKGVE---
sp|P02626|PRVA_AMPME   LKGFSKEGRELTDKETKDLLIKGDKDGDGKIGVDEFTSLVAES----
sp|P20472|PRVA_HUMAN   LKGFSPDARDLSAKETKMLMAAGDKDGDGKIGVDEFSTLVAES----
sp|P80079|PRVA_FELCA   LKGFYPDARDLSVKETKMLMAAGDKDGDGKIDVDEFFSLVAKS----
sp|P32930|ONCO_HUMAN   LQKFESGARELTESETKSLMAAADNDGDGKIGAEEFQEMVHS-----
sp|P43305|PRVU_CHICK   LQRFECGARVLTASETKTFLAAADHDGDGKIGAEEFQEMVQS-----
sp|P02619|PRVB_ESOLU   LQNFSPSARALTDAETKAFLADGDKDGDGMIGVDEFAAMIKA-----
sp|Q91482|PRVB1_SALSA  LQNFCPKARELTDAETKAFLKAGDADGDGMIGIDEFAVLVKQ-----
sp|P02620|PRVB_MERME   LQNFSAGARALTDAETATFLKAGDSDGDGKIGVEEFAAMVKG-----
sp|P02622|PRVB_GADCA   LIAFAADLRALTDAETKAFLKAGDSDGDGKIGVDEFGALVDKWGAKG
                       :       . ::  :   ::  .* :.** *. :**  ::
```

Figure 9-11: Including more distant relatives in the multiple alignment.

With such a result, we can safely bet that these two conserved regions are surely involved in some biological function. We know that most of these proteins bind calcium, so we can safely bet that the calcium-binding site involves some of these conserved positions. This is indeed the case as revealed by the Swiss-Prot annotation.

Comparing Sequences That You Can't Align

Sometimes you need to compare sequences that don't necessarily have a common ancestor — or are so distantly related that considering them as homologous is difficult.

Multiple-sequence-alignment programs are usually hopeless in these situations. They'll include proteins that contain homologous domains but that are otherwise unrelated, or segments of DNA that seem to contain similar regulating features although it is impossible to see anything common between them in a multiple alignment.

When a situation like this arises, do not expect miracles from bioinformatics. In most cases, your first intuition that there is nothing to get out of these sequences proves to be the right one. Nonetheless, before giving up completely, it is worth trying the two techniques we briefly introduce in this section. The first one, the Gibbs sampler, simultaneously looks into all your sequences for short, partially conserved gap-free segments. Pratt, the second method, looks for flexible patterns — a special category of segments that can contain gaps, and that need to be conserved only at certain positions.

Making multiple local alignments with the Gibbs sampler

The Gibbs sampler is a stochastic method: It first scrambles your sequences, aligns them randomly until a good solution appears, and then keeps scrambling the sequences to improve this good solution. The interesting thing about using chance is that you can solve complicated problems without having to explore all the possible solutions.

Gurus name this type of method *stochastic* because it contains an element of chance. Stochastic methods are by far the most powerful in bioinformatics. Unfortunately, they do not have the popularity they deserve because they make it difficult to reproduce the same results twice. For instance, if you run the Gibbs sampler twice on the same set of sequences, you may not get exactly the same solution. For most biologists, reproducibility is paramount, and — let's face it — we all hate it when computers change their minds from one minute to the next.

Nonetheless, if you're ready for a bit of fuzzy logic, you may find that the Gibbs sampler can offer very sensible solutions to extremely complicated problems. For instance, the Gibbs sampler is very good at identifying HTH (Helix-Turn-Helix) domains across a protein family. It is also a nice way to search for regulatory elements shared by otherwise unrelated DNA sequences. You can access the Gibbs sampler from its home page at

```
bayesweb.wadsworth.org/gibbs/gibbs.html
```

or use the server the Pasteur Institute maintains at

```
bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html
```

When using a Gibbs sampler, bear in mind that to be accurate, it needs as many sequences as possible. You should not use it with less than 20 sequences.

Searching conserved patterns

The Gibbs sampler is useful only when the segments you're looking for have exactly the same length, like Helix-Turn-Helix domains. However, the vast majority of the interesting motifs you may come across when doing bioinformatics have different lengths; the most standard situation is to find motifs that are poorly conserved. When this happens, your only hope is to find a few highly conserved amino acids or nucleotides that anchor the motif.

If your sequences are poorly related but contain such motifs, then the only way to analyze them is to use a pattern-finding motif. If you want to know more about patterns (biologists often call them PROSITE patterns), you can go to Chapter 7. Several tools exist for identifying these conserved patterns in a set of unaligned sequences. Showing you how to use these tools and to interpret their result is beyond the scope of this book. However, if you want to experiment with them, you should know that Pratt is one of the most powerful because it allows some flexible spacing between the conserved positions. You can also use TEIRESIAS, MEME, or SMILE. (Flip ahead to Table 9-10 to see a partial list of these resources.)

Internet Resources for Doing Multiple Sequence Comparisons

The amount of resources for making multiple alignments online is almost overwhelming. The usual words of caution apply to the resources that we list in the following tables:

- Use stable resources and always make a few simple tests to make sure that the service you're using does what it's supposed to do.
- Never depend blindly on a resource you don't control. As they say, "Good servers go bad, and bad servers go down."
- If you want to do many multiple sequence alignments — or if your company does not authorize you to send your sequences over the Internet — you may have to install these programs on your own machine. We give you addresses where you can download the source code or the executable files of some of these programs.

Making multiple alignments with ClustalW around the clock

ClustalW is the most common online program. Table 9-8 gives you a list of available servers, starting with the site where you can download ClustalW to install it on your own machine.

Note: The last link in Table 9-8 is the location where you can download the executable to install it on your own machine. It is *not* a Web server.

Table 9-8 A List of ClustalW Servers

Name	*Location*	*Address*
EBI	Europe	`www.ebi.ac.uk/clustalw`
EMBnet	Europe	`www.ch.embnet.org/software/ClustalW.html`
PIR	USA	`pir.georgetown.edu/pirwww/search/multialn.shtml`
BCM	USA	`searchlauncher.bcm.tmc.edu/multi-align/multi-align.html`
GenomeNet	Japan	`align.genome.jp/`
DDBJ	Japan	`www.ddbj.nig.ac.jp/search/clustalw-e.html`
Strasbourg	Europe	`ftp://ftp-igbmc.u-strasbg.fr/pub/ClustalW/`

Finding your favorite alignment method

If ClustalW isn't suited for the sequences you want to align, Table 9-9 saves the day by giving you the addresses of a few other servers.

Table 9-9 Multiple Sequence Alignment Resources over the Internet

Method	*Description*	*Address*

Tcoffee	Accurate combination of sequences and structures	`www.tcoffee.org`
		`www.ch.embnet.org/software/TCoffee.html`
		`www.ebi.ac.uk/t-coffee/`

Probcons	A Bayesian version of Tcoffee	`probcons.stanford.edu/`
MUSCLE	A fast and accurate sequence cruncher	`www.drive5.com/muscle/`
Kalign	A fast sequence aligner	`msa.cgb.ki.se`
MAFFT	A fast and accurate sequence cruncher using Fast Fourier Transforms	`timpani.genome.ad.jp/~mafft/server/`
Dialign	Ideal for Sequences With Local Homology	`bibiserv.techfak.uni-bielefeld.de/dialign/`

Searching for motifs or patterns

If your sequences are too distantly related for making a multiple sequence alignment, your last chance is trying to find a conserved pattern across them. Table 9-10 lists some of the online methods you can use for this purpose.

Table 9-10 Motif-finding Methods Available Online

Method	*Address*
Gibbs Sampler	`bioweb.pasteur.fr/seqanal/interfaces/gibbs-simple.html`, `bayesweb.wadsworth.org/gibbs/gibbs.html`
Pratt	`http://www.ebi.ac.uk/pratt/`
eMotif	`dna.stanford.edu/emotif/`

(continued)

Table 9-10 ***(continued)***

Method	*Address*
MEME	`http://meme.sdsc.edu/meme/`
TEIRESIAS	`http://cbcsrv.watson.ibm.com/Tspd.html`
Bioprospector	`http://ai.stanford.edu/~xsliu/BioProspector/`
Improbizer	`www.soe.ucsc.edu/~kent/improbizer/improbizer.html`
BLOCK-Maker	`blocks.fhcrc.org/blocks/blockmkr/make_blocks.html`

Chapter 10

Editing and Publishing Alignments

In This Chapter

- Reformatting your multiple sequence alignment
- Manually editing a multiple sequence alignment with Jalview
- Turning your Alignment into a logo
- Shading a multiple sequence alignment with Boxshade

> *It looked so good that I thought it just had to be genuine.*
>
> — Everyone's favorite secret thought

This chapter is all about using and modifying an existing multiple sequence alignment. If you don't have one ready yet, don't worry; we start by showing you how to generate a dummy multiple sequence alignment that you can use throughout the entire chapter.

The first real section in this chapter — as opposed to the kind-of-real section about making a dummy multiple sequence alignment — tells you what gurus assume we all know about alignment formats. We also show you how to decide which format you want to use (or avoid!) and — generally speaking — how to find your way across the format jungle.

If you have generated your multiple sequence alignment with any kind of automatic method, you probably need to arrange (edit) it manually before you can really use it. When you stop to think about it, editing a multiple sequence alignment isn't trivial: It requires inserting gaps in an entire subgroup of sequences or shifting several sequences all at once, and so on. When you do this with a standard word processor, pretty soon you're going to feel as if you're playing with a Rubik's Cube. No matter how slowly and carefully you start, after three minutes of moving things about, everything is out of control! In order to avoid this kind of confusion, we show you some powerful editing tools in this chapter that you can use over the Internet.

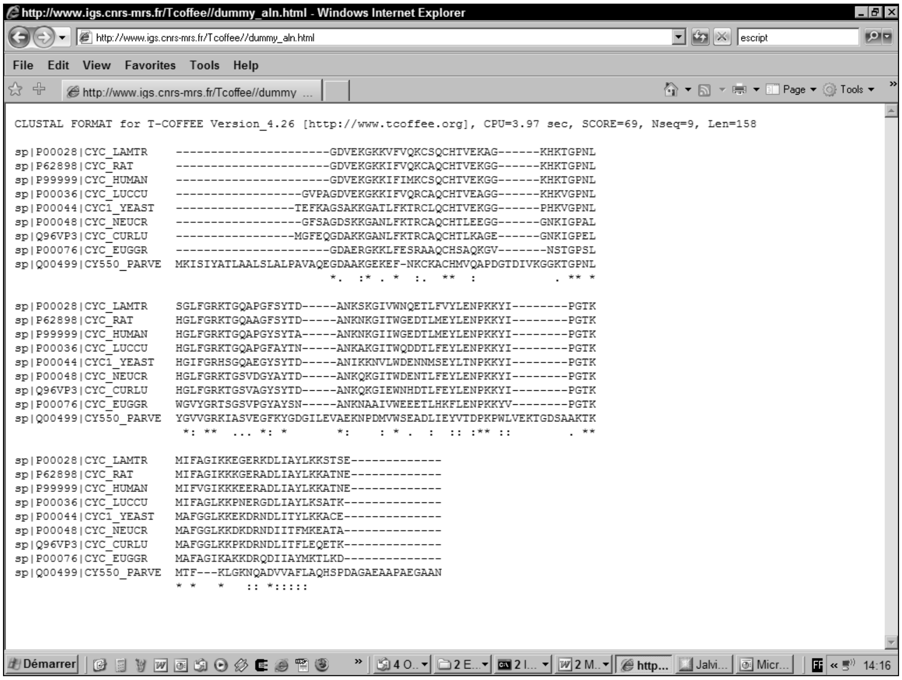
After you have the alignment of your dreams, there's a good chance that you may want to prepare it for publication. Nothing looks nicer than a beautiful multiple alignment with a well-chosen color scheme and a few pieces of annotation where it matters. In fact, good-looking alignments are so convincing that you should convince yourself of the quality of your alignment *before* beautifying it! Several online tools are available to do this, and we introduce them in this chapter.

Finally, we show you a couple of tools that can help you extract information from your multiple sequence alignment. These include using protein logos to identify the most conserved positions.

All in all, this is a small chapter — but if you decide to go through it, you'll have no more excuses for attaching simple, raw-text alignments to your reports, publications, or lab books!

TIP

If you do not have an alignment ready, you can download one for your crash tests from `www.tcoffee.org/dummy_aln.html`. It is an alignment in ALN format (see Figure 10-1).


http://www.igs.cnrs-mrs.fr/Tcoffee//dummy_aln.html - Windows Internet Explorer
http://www.igs.cnrs-mrs.fr/Tcoffee//dummy_aln.html
escript
File Edit View Favorites Tools Help
http://www.igs.cnrs-mrs.fr/Tcoffee//dummy ...
Page
Tools
Démarrer
4 O...
2 E...
2 I...
2 M...
http...
Jalvi...
Micr...
14:16

```
CLUSTAL FORMAT for T-COFFEE Version_4.26 [http://www.tcoffee.org], CPU=3.97 sec, SCORE=69, Nseq=9, Len=158

sp|P00028|CYC_LAMTR    ----------------------GDVEKGKKVFVQKCSQCHTVEKAG------KHKTGPNL
sp|P62898|CYC_RAT      ----------------------GDVEKGKKIFVQKCAQCHTVEKGG------KHKTGPNL
sp|P99999|CYC_HUMAN    ----------------------GDVEKGKKIFIMKCSQCHTVEKGG------KHKTGPNL
sp|P00036|CYC_LUCCU    ------------------GVPAGDVEKGKKIFVQRCAQCHTVEAGG------KHKVGPNL
sp|P00044|CYC1_YEAST   -----------------TEFKAGSAKKGATLFKTRCLQCHTVEKGG------PHKVGPNL
sp|P00048|CYC_NEUCR    ------------------GFSAGDSKKGANLFKTRCAQCHTLEEGG------GNKIGPAL
sp|Q96VP3|CYC_CURLU    -----------------MGFEQGDAKKGANLFKTRCAQCHTLKAGE------GNKIGPEL
sp|P00076|CYC_EUGGR    ----------------------GDAERGKKLFESRAAQCHSAQKGV-------NSTGPSL
sp|Q00499|CY550_PARVE  MKISIYATLAALSLALPAVAQEGDAAKGEKEF-NKCKACHMVQAPDGTDIVKGGKTGPNL
                                             *.  :* . *  :.  **  :           . ** *

sp|P00028|CYC_LAMTR    SGLFGRKTGQAPGFSYTD-----ANKSKGIVWNQETLFVYLENPKKYI--------PGTK
sp|P62898|CYC_RAT      HGLFGRKTGQAAGFSYTD-----ANKNKGITWGEDTLMEYLENPKKYI--------PGTK
sp|P99999|CYC_HUMAN    HGLFGRKTGQAPGYSYTA-----ANKNKGIIWGEDTLMEYLENPKKYI--------PGTK
sp|P00036|CYC_LUCCU    HGLFGRKTGQAPGFAYTN-----ANKAKGITWQDDTLFEYLENPKKYI--------PGTK
sp|P00044|CYC1_YEAST   HGIFGRHSGQAEGYSYTD-----ANIKKNVLWDENNMSEYLTNPKKYI--------PGTK
sp|P00048|CYC_NEUCR    HGLFGRKTGSVDGYAYTD-----ANKQKGITWDENTLFEYLENPKKYI--------PGTK
sp|Q96VP3|CYC_CURLU    HGLFGRKTGSVAGYSYTD-----ANKQKGIEWNHDTLFEYLENPKKYI--------PGTK
sp|P00076|CYC_EUGGR    WGVYGRTSGSVPGYAYSN-----ANKNAAIVWEEETLHKFLENPKKYV--------PGTK
sp|Q00499|CY550_PARVE  YGVVGRKIASVEGFKYGDGILEVAEKNPDMVWSEADLIEYVTDPKPWLVEKTGDSAAKTK
                        *: **  ... *: *       *:    : * .  :  :: :** ::        .  **

sp|P00028|CYC_LAMTR    MIFAGIKKEGERKDLIAYLKKSTSE-------------
sp|P62898|CYC_RAT      MIFAGIKKKGERADLIAYLKKATNE-------------
sp|P99999|CYC_HUMAN    MIFVGIKKKEERADLIAYLKKATNE-------------
sp|P00036|CYC_LUCCU    MIFAGLKKPNERGDLIAYLKSATK--------------
sp|P00044|CYC1_YEAST   MAFGGLKKEKDRNDLITYLKKACE--------------
sp|P00048|CYC_NEUCR    MAFGGLKKDKDRNDIITFMKEATA--------------
sp|Q96VP3|CYC_CURLU    MAFGGLKKPKDRNDLITFLEQETK--------------
sp|P00076|CYC_EUGGR    MAFAGIKAKKDRQDIIAYMKTLKD--------------
sp|Q00499|CY550_PARVE  MTF---KLGKNQADVVAFLAQHSPDAGAEAAPAEGAAN
                       * *   *   :: *:::::
```

Figure 10-1: A multiple sequence alignment in ALN or ClustalW format.

Getting Your Multiple Alignment in the Right Format

When using online servers — or even local programs — you do not always have full control of what goes in and what comes out of the program you use. For instance, many multiple-sequence-alignment programs commonly output only one format: MSF (Multiple Sequence Format). What can you do if you want to analyze this multiple sequence alignment with a program that only reads FASTA-formatted alignments? The answer is simple: Use a reformatting program just like the ones we introduce in this chapter.

When it comes to multiple sequence alignments, there is no such thing as a perfect format. Anyone used to dealing with the many services available online knows that each service tends to have its favorite format. (See the "A format for everyone: Democracy or anarchy?" sidebar if you want proof of this fact.) Understanding everything there is to know about formats is an ability you only acquire with some practice — but this chapter gives you some ideas, and we prepare you to be on the lookout for potential problems.

A format for everyone: Democracy or anarchy?

The variety of formats is a major curse in bioinformatics. For a long time, the tradition was that anyone developing a new program would design his or her own house format. Each new format was slightly different from the others — and especially appealing to only a particular category of biologists. For instance, people doing phylogeny became very attached to the Phylip format; users of GCG (a popular bioinformatics package) preferred MSF, and so on. Whether we like it or not, the weight of history has made each of these formats totally acceptable — and today they all live in perfect harmony, side by side on our computer keyboards — or do they?

Another source of confusion in the field was the communication wars between biologists and computer scientists. For many years, biologists have complained about the computer scientists' formats, basically saying, "We cannot make sense of this gibberish." In retaliation (and because they needed to), biologists also created formats they could use — to the contempt of computer scientists, who dubbed these formats "amateurish and ambiguous." Both the biologists and the computer scientists were probably right in their respective evaluations. But that didn't help.

Things may be about to change with the XML language coming to the fore. XML (e*X*tensible *M*arkup *L*anguage) is a close relative of HTML — the language of the Web. XML makes it possible to simply describe your data with keywords everyone can agree on. Today, biologists and computer scientists consent — albeit weakly — that XML could be the solution and the main bioinformatics programs are now able to produce output in XML.

Unfortunately, nothing is standardized yet, and the number of programs able to use XML is still limited. To be honest, you can expect the ancient formats we describe in this section to still be around when your grandchildren retire in their early 90s.

When you choose to store your data in a specific format, you must ask yourself four questions:

- Do most programs support this format?
- Will my collaborators be able to use it?
- Can I store all the information I need with this format?
- Is it easy to manipulate?

Today no format is so perfect that you would answer yes to each of these questions; any choice relies on the right trade-off between what you really need and what's available. If you can, avoid using a mixture of formats when running a project.

The taxonomy of multiple-sequence-alignment formats is rather complex. Table 10-1 lists the most common formats you may come across when manipulating multiple sequence alignments.

Table 10-1 A Classification of Multiple Sequence Alignment Formats

Name	*Type*	*Usage*
post-script, PDF, HTML	Graphic	Terminal formats suitable for printing only
FASTA (Figure 10-2)	Text	Easy to manipulate Noninterleaved (not readable by humans) Easy for renaming sequences Supported by most programs Cannot incorporate extra annotation
PIR	Text	Similar to FASTA but with an extra line Can incorporate limited extra annotation
MSF (Figure 10-3)	Text	Most standard multiple-alignment format Difficult to modify manually Interleaved (easy to read for humans) Can include extra annotation, such as weights Supported by most programs (but some programs only support it partially!)

Name	Type	Usage
Selex	Text	Extended version of MSF Can include extra annotation Supported by few programs
ALN (Figure 10-1)	Text	Simplified version of MSF Default output of ClustalW Interleaved (easy to read for humans) Supported by many programs
Phylip	Text	Variant of ALN Useful for doing phylogenetic analysis Supported by most phylogenetic packages

Recognizing the main formats

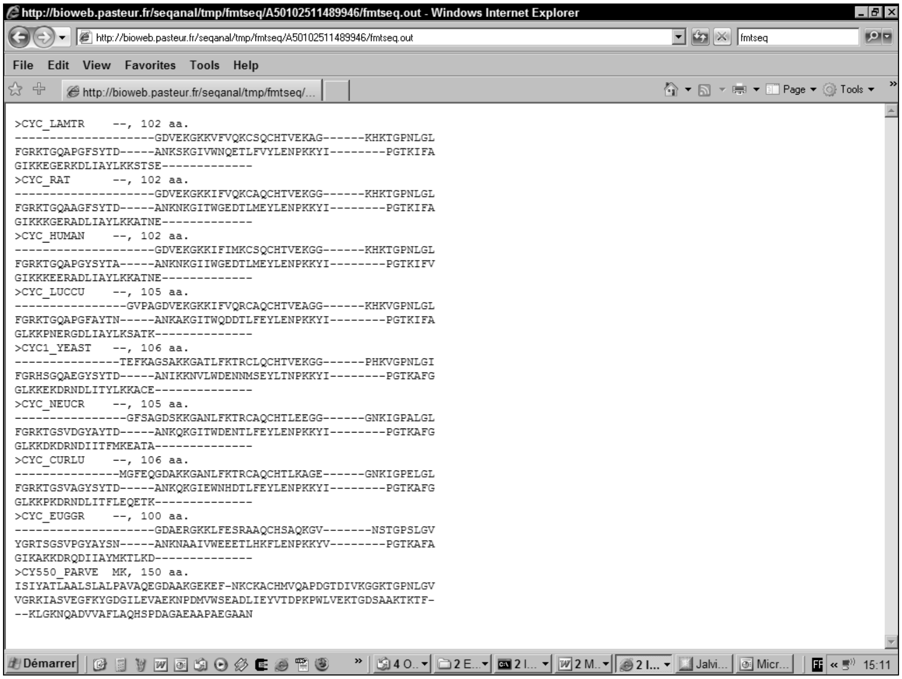
The alignments you see in publications are usually in a graphic format that isn't very useful for doing analyses. The last section of this chapter shows you resources that you can use for generating these alignments. You've already seen the ALN format in this chapter (refer to Figure 10-1), and FASTA is another format (see Figure 10-2) that has cropped up a number of times in previous chapters. In the lab, MSF, as shown in Figure 10-3, is probably the most common format. Note that MSF and ALN are *interleaved* (meaning easier for humans to read but not for machines), whereas FASTA is not.

To make an interleaved format, the idea is to take the complete multiple sequence alignment — and then chop it into blocks of 60 residues. The file displays the blocks one after the other.

If you want to further investigate the intricacies of the world of formatting, here are two extensive resources for keeping your formats straight: `emboss.sourceforge.net/docs/themes/SequenceFormats.html` and `www.ebi.ac.uk/help/formats_frame.html`.

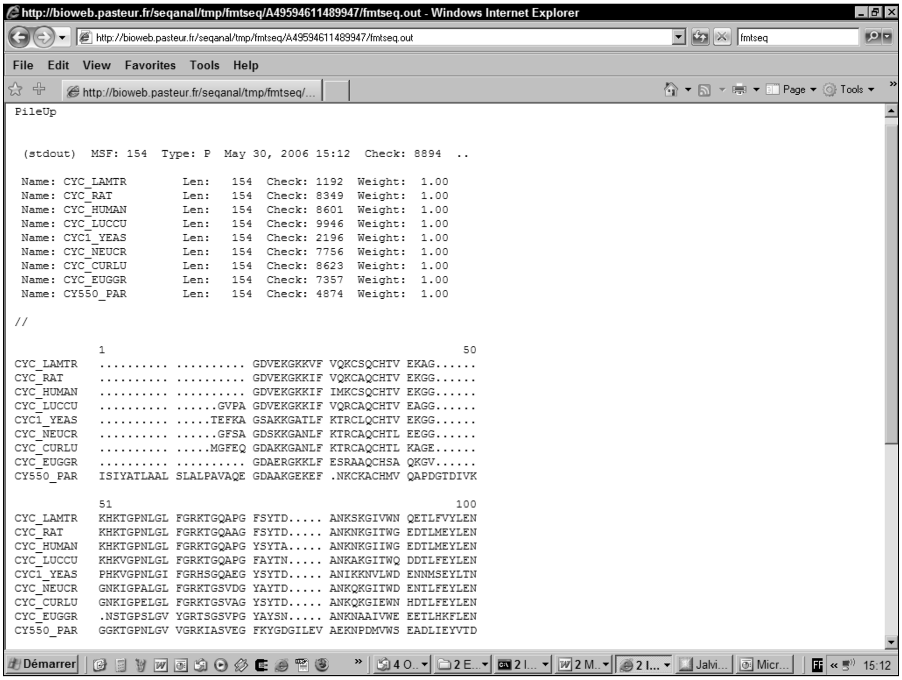
Working with the right format

Sometimes when a program offers several outputs, keeping only the graphic output is tempting because it seems — with its flashy colored annotation — to be the most complete. This graphic output usually comes in post-script, `.pdf`, `.html`, `.gif`, or `.jpeg` format. If you keep only this graphic output, though, you can be badly stuck.


http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/A50102511489946/fmtseq.out - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/A50102511489946/fmtseq.out
fmtseq
File Edit View Favorites Tools Help
http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/...
Page
Tools
Démarrer
4 O...
2 E...
2 I...
2 M...
2 I...
Jalvi...
Micr...
15:11

```
>CYC_LAMTR    --, 102 aa.
--------------------GDVEKGKKVFVQKCSQCHTVEKAG------KHKTGPNLGL
FGRKTGQAPGFSYTD-----ANKSKGIVWNQETLFVYLENPKKYI--------PGTKIFA
GIKKEGERKDLIAYLKKSTSE-------------
>CYC_RAT      --, 102 aa.
--------------------GDVEKGKKIFVQKCAQCHTVEKGG------KHKTGPNLGL
FGRKTGQAAGFSYTD-----ANKNKGITWGEDTLMEYLENPKKYI--------PGTKIFA
GIKKKGERADLIAYLKKATNE-------------
>CYC_HUMAN    --, 102 aa.
--------------------GDVEKGKKIFIMKCSQCHTVEKGG------KHKTGPNLGL
FGRKTGQAPGYSYTA-----ANKNKGIIWGEDTLMEYLENPKKYI--------PGTKIFV
GIKKKEERADLIAYLKKATNE-------------
>CYC_LUCCU    --, 105 aa.
----------------GVPAGDVEKGKKIFVQRCAQCHTVEAGG------KHKVGPNLGL
FGRKTGQAPGFAYTN-----ANKAKGITWQDDTLFEYLENPKKYI--------PGTKIFA
GLKKPNERGDLIAYLKSATK--------------
>CYC1_YEAST   --, 106 aa.
---------------TEFKAGSAKKGATLFKTRCLQCHTVEKGG------PHKVGPNLGI
FGRHSGQAEGYSYTD-----ANIKKNVLWDENNMSEYLTNPKKYI--------PGTKAFG
GLKKEKDRNDLITYLKKACE--------------
>CYC_NEUCR    --, 105 aa.
----------------GFSAGDSKKGANLFKTRCAQCHTLEEGG------GNKIGPALGL
FGRKTGSVDGYAYTD-----ANKQKGITWDENTLFEYLENPKKYI--------PGTKAFG
GLKKDKDRNDIITFMKEATA--------------
>CYC_CURLU    --, 106 aa.
---------------MGFEQGDAKKGANLFKTRCAQCHTLKAGE------GNKIGPELGL
FGRKTGSVAGYSYTD-----ANKQKGIEWNHDTLFEYLENPKKYI--------PGTKAFG
GLKKPKDRNDLITFLEQETK--------------
>CYC_EUGGR    --, 100 aa.
--------------------GDAERGKKLFESRAAQCHSAQKGV-------NSTGPSLGV
YGRTSGSVPGYAYSN-----ANKNAAIVWEEETLHKFLENPKKYV--------PGTKAFA
GIKAKKDRQDIIAYMKTLKD--------------
>CY550_PARVE  MK, 150 aa.
ISIYATLAALSLALPAVAQEGDAAKGEKEF-NKCKACHMVQAPDGTDIVKGGKTGPNLGV
VGRKIASVEGFKYGDGILEVAEKNPDMVWSEADLIEYVTDPKPWLVEKTGDSAAKTKTF-
--KLGKNQADVVAFLAQHSPDAGAEAAPAEGAAN
```

Figure 10-2: Multiple sequence alignment in FASTA format.


http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/A49594611489947/fmtseq.out - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/A49594611489947/fmtseq.out
fmtseq
File Edit View Favorites Tools Help
http://bioweb.pasteur.fr/seqanal/tmp/fmtseq/...
Page
Tools
Démarrer
4 O...
2 E...
2 I...
2 M...
2 I...
Jalvi...
Micr...
15:12

```
PileUp


 (stdout)  MSF: 154  Type: P  May 30, 2006 15:12  Check: 8894  ..

 Name: CYC_LAMTR        Len:   154  Check: 1192  Weight:  1.00
 Name: CYC_RAT          Len:   154  Check: 8349  Weight:  1.00
 Name: CYC_HUMAN        Len:   154  Check: 8601  Weight:  1.00
 Name: CYC_LUCCU        Len:   154  Check: 9946  Weight:  1.00
 Name: CYC1_YEAS        Len:   154  Check: 2196  Weight:  1.00
 Name: CYC_NEUCR        Len:   154  Check: 7756  Weight:  1.00
 Name: CYC_CURLU        Len:   154  Check: 8623  Weight:  1.00
 Name: CYC_EUGGR        Len:   154  Check: 7357  Weight:  1.00
 Name: CY550_PAR        Len:   154  Check: 4874  Weight:  1.00

//

            1                                                50
CYC_LAMTR   .......... .......... GDVEKGKKVF VQKCSQCHTV EKAG......
CYC_RAT     .......... .......... GDVEKGKKIF VQKCAQCHTV EKGG......
CYC_HUMAN   .......... .......... GDVEKGKKIF IMKCSQCHTV EKGG......
CYC_LUCCU   .......... ......GVPA GDVEKGKKIF VQRCAQCHTV EAGG......
CYC1_YEAS   .......... .....TEFKA GSAKKGATLF KTRCLQCHTV EKGG......
CYC_NEUCR   .......... ......GFSA GDSKKGANLF KTRCAQCHTL EEGG......
CYC_CURLU   .......... .....MGFEQ GDAKKGANLF KTRCAQCHTL KAGE......
CYC_EUGGR   .......... .......... GDAERGKKLF ESRAAQCHSA QKGV......
CY550_PAR   ISIYATLAAL SLALPAVAQE GDAAKGEKEF .NKCKACHMV QAPDGTDIVK

            51                                              100
CYC_LAMTR   KHKTGPNLGL FGRKTGQAPG FSYTD..... ANKSKGIVWN QETLFVYLEN
CYC_RAT     KHKTGPNLGL FGRKTGQAAG FSYTD..... ANKNKGITWG EDTLMEYLEN
CYC_HUMAN   KHKTGPNLGL FGRKTGQAPG YSYTA..... ANKNKGIIWG EDTLMEYLEN
CYC_LUCCU   KHKVGPNLGL FGRKTGQAPG FAYTN..... ANKAKGITWQ DDTLFEYLEN
CYC1_YEAS   PHKVGPNLGI FGRHSGQAEG YSYTD..... ANIKKNVLWD ENNMSEYLTN
CYC_NEUCR   GNKIGPALGL FGRKTGSVDG YAYTD..... ANKQKGITWD ENTLFEYLEN
CYC_CURLU   GNKIGPELGL FGRKTGSVAG YSYTD..... ANKQKGIEWN HDTLFEYLEN
CYC_EUGGR   .NSTGPSLGV YGRTSGSVPG YAYSN..... ANKNAAIVWE EETLHKFLEN
CY550_PAR   GGKTGPNLGV VGRKIASVEG FKYGDGILEV AEKNPDMVWS EADLIEYVTD
```

Figure 10-3: A multiple sequence alignment in MSF format.

Graphic outputs are useful for displaying your information. Using graphic output with any sequence-analysis program, however, is virtually impossible. You can request a graphic output if you want the pretty picture, but you must also keep the text version. Remember that turning a text into a graphic is always easier than the other way round!

Text representations are much more civilized; they are the default output of most multiple-sequence-alignment programs and servers. Their main purpose is to be easy to display, easy to manipulate, and easy to transfer from one program to the next. Of course, they're much less flashy than their graphic counterparts!

MSF is the most standard format for multiple sequence alignments. This is the most common format if you download an alignment from the Web, for example. FASTA or PIR are a bit less common but are on the rise. If you're using phylogenetic programs, Phylip is fairly common. ALN is also fairly common because it is the default output of ClustalW, the most widely used multiple-sequence-alignment program. ALN offers a good trade-off between most formats.

Converting formats

If the program you're using doesn't produce alignments in the format you need, it is possible to use a third-party conversion tool to get to the format you want. In the following step list, we show you how to convert an alignment in ALN format (refer to Figure 10-1) into a FASTA-format multiple alignment.

1. **Point your browser to** `bioweb.pasteur.fr/seqanal/interfaces/fmtseq.html`**.**

   The Pasteur Institute's fmtseq Sequence Conversion page appears, as shown in Figure 10-4.

2. **Enter your e-mail address in the field provided.**

3. **Paste your alignment in the Sequence window.**

   fmtseq automatically recognizes most formats. If you don't have any spare sequences lying around, use the dummy multiple sequence alignment we created at the beginning of the chapter.

   If your sequence names are longer than 15 characters, the rest of the name gets incorporated IN the sequence. This is a major source of confusion.

4. **Specify a format for your output by making a selection from the Output Sequence Format drop-down menu.**

   For our example, we chose FASTA.

5. **Scroll Down The Page to reach the "Input Parameters" section (you can also click on the Input Parameters link) and select the format you want to input from the Input Sequence Format drop-down menu.**

If you leave this selector blank, as shown in Figure 10-5, fmtseq tries to guess your format automatically. If it fails to do so, you can select your input format here (assuming you know it!).

6. **In the Output Parameters section, make sure that the Convert to Lowercase and the Convert to Uppercase options are deselected.**

   When you deselect these two options, fmtseq retains the original case of each symbol in your alignment. Some programs, such as Dialign, use case to indicate the local reliability of the alignment.

   The Pretty-Print option gives you access to a menu that makes it possible to control every detail in your output format, such as the number of characters per line, the indentation, and other minute details.

7. **Click the Run ftmseq button at the top or the bottom of the page.**

   Wait until your browser displays the result page. In most cases, the results come back interactively in a few seconds, but if the server is very busy, it may send the output using the e-mail you provided in Step 2.

8. **Click the fmtseq.out link to see your alignment reformatted.**

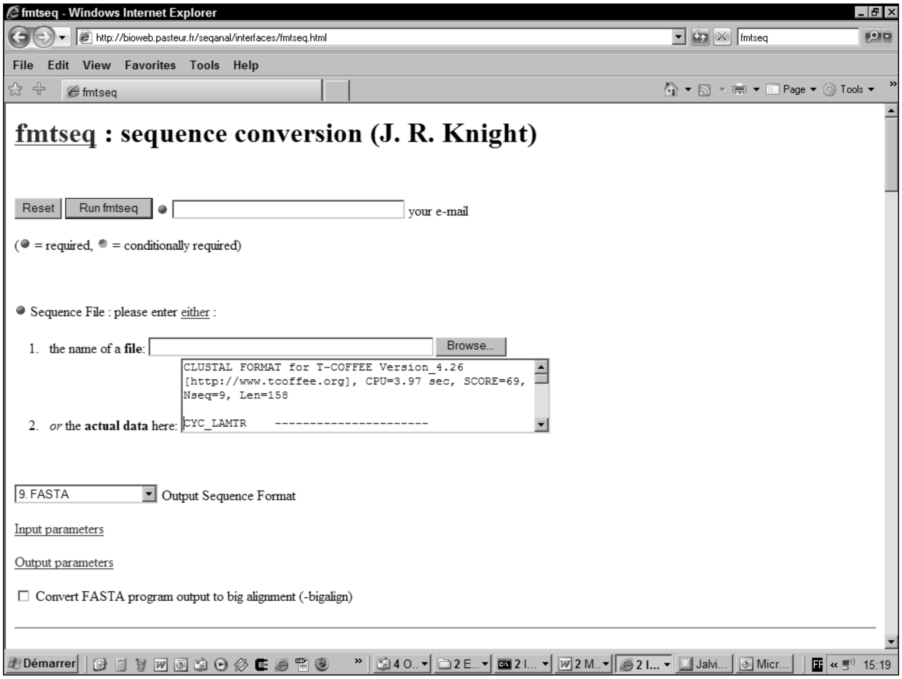
9. **Save the reformatted file with the File➪Save As option of your browser.**

   You can save this file as a `.txt` or `.html` file.

Sequence reformatting is one of the most common facilities on the Web. Table 10-2 lists some of these sequence reformatting servers and what you can do using them.

Table 10-2 Sequence Text Conversion on the Web

Name	*Address*	*Description*
fmtseq	`bioweb.pasteur.fr/seqanal/interfaces/fmtseq-simple.html`, `www.bimcore.emory.edu/Pise/`	Converts most formats
READSEQ	`bimas.dcrt.nih.gov/molbio/readseq/`, `dot.imgen.bcm.tmc.edu/seq-util/readseq.html`, `iubio.bio.indiana.edu/cgi-bin/readseq.cgi`, `bimas.dcrt.nih.gov/molbio/readseq/`, `www.ebi.ac.uk/readseq/`	A very popular tool for reformatting; does not recognize ALN
SeqCheck	`darwin.nmsu.edu/bioinfo/seqcheck/seqcheck.php`	Cleans your FASTA sequences


fmtseq - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/interfaces/fmtseq.html
fmtseq
File Edit View Favorites Tools Help
fmtseq
Page ▾ Tools ▾
Démarrer
15:19

fmtseq : sequence conversion (J. R. Knight)

Reset | Run fmtseq | ● [ ] your e-mail

(● = required, ● = conditionally required)

● Sequence File : please enter either :

1. the name of a file: [ ] Browse...
2. *or* the actual data here:

```
CLUSTAL FORMAT for T-COFFEE Version_4.26
[http://www.tcoffee.org], CPU=3.97 sec, SCORE=69,
Nseq=9, Len=158

CYC_LAMTR    ----------------------
```

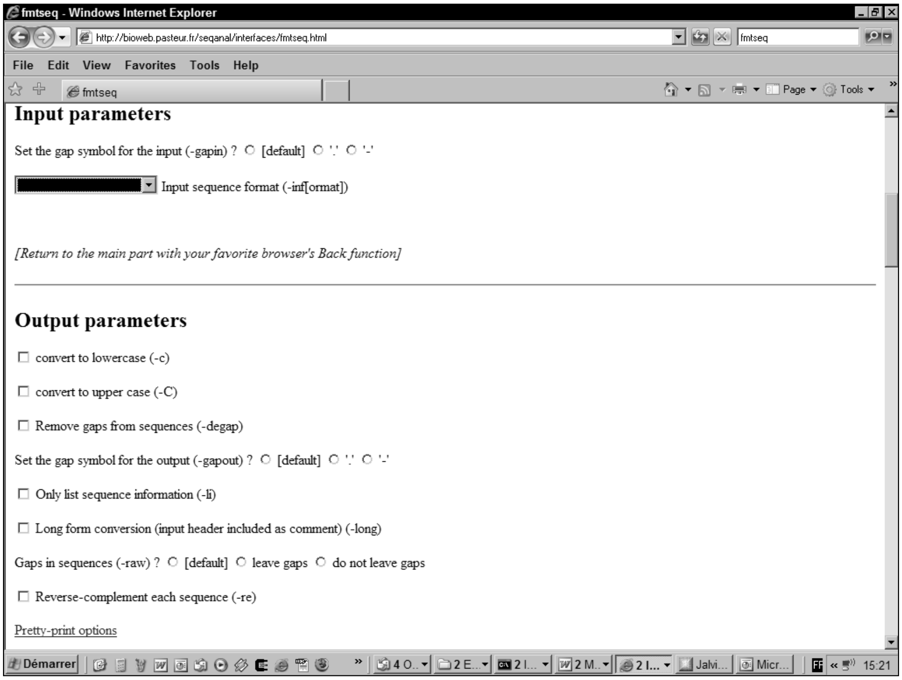
9. FASTA ▾ Output Sequence Format

Input parameters

Output parameters

□ Convert FASTA program output to big alignment (-bigalign)

Figure 10-4: The fmtseq server.


fmtseq - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/interfaces/fmtseq.html
fmtseq
File Edit View Favorites Tools Help
fmtseq
Page ▾ Tools ▾
Démarrer
15:21

Input parameters

Set the gap symbol for the input (-gapin) ? ○ [default] ○ '.' ○ '-'

[ ] ▾ Input sequence format (-inf[ormat])

[Return to the main part with your favorite browser's Back function]

Output parameters

□ convert to lowercase (-c)

□ convert to upper case (-C)

□ Remove gaps from sequences (-degap)

Set the gap symbol for the output (-gapout) ? ○ [default] ○ '.' ○ '-'

□ Only list sequence information (-li)

□ Long form conversion (input header included as comment) (-long)

Gaps in sequences (-raw) ? ○ [default] ○ leave gaps ○ do not leave gaps

□ Reverse-complement each sequence (-re)

Pretty-print options

Figure 10-5: Output parameters of fmtseq.

Watching out for lost data

No matter what they tell you, there's always the danger that you can lose information when changing the format of your sequences. Each type of format supports different types of information. The problem is that, in most cases, you only realize this when it's too late. Table 10-3 lists the kinds of features that formatting can destroy.

Along the same lines, do not take for granted that similar online servers do the same thing (even if they have the same name and the same interface). Two servers running READSEQ may run different versions of this program, or the same version with different default parameters. As a consequence, a problem that doesn't occur with one server may occur with the next server. It really pays to keep your eyes peeled and to keep backup copies of your original files.

Table 10-3 Information You Can Lose When Reformatting

Information Type	*Nature of the Loss*
Sequence name	Long names can be truncated when switching formats. Special characters may also be modified. This happens when converting from FASTA to ALN. The effect of the truncation is unpredictable. Sometimes a portion of the name is added to the sequence!
Upper/lowercase	Case sometimes contains useful information. Strictly speaking, FASTA only supports uppercase. Some programs hate to receive an input with a mixture of cases.
Gap type	Some formats, such as MSF, use different symbols for different types of gaps (. , –, ~). These are often turned into (–) symbols after reformatting.
Annotation	MSF can support weight values for the sequences. This information is lost in most conversions. The extra line of annotation in PIR often disappears when changing the format of these alignments. Any annotation that comes after the sequence name in FASTA is bound to disappear after a conversion.
Special amino acid or nucleotides	Some reformatting programs support the code for ambiguities such as X (for undetermined amino acids) or N (for nucleotides). If your sequences travel across many programs, these symbols may disappear. This can be a problem if you rely on the offset of some residues within the sequences. If you can, stick to the standard 4-nucleotide alphabet and the 20-amino-acid alphabet.

Using Jalview to Edit Your Multiple Alignment Online

Multiple-sequence-alignment methods aren't perfect. The people who make them know this, and the people who use them must also consider this fact.

To produce an alignment that you know is right, you can torture a program (and yourself) until it spits out exactly what you want. This may take time and may never lead to the desired result. You can also bite the bullet and edit this alignment yourself. The best recipe for instant insanity is trying to modify your multiple sequence alignment with a standard word processor (such as Microsoft Word or any similar product). It is an experience that has shattered some of the smartest people. The chances of never recovering are high, so don't do it!

Because editing a multiple sequence alignment is so complicated, biologists have developed text editors that are specific for multiple sequence alignment. They make it easy for you to see exactly what's going on. They also make it possible to group sequences so you can realign two subgroups without having to modify the entire alignment. Most of these editors require that you install something on your computer. This should not deter you; some of these programs — Seaview, for instance — are really easy to install on a PC.

However, if you want to stick to your browser, you can use Jalview, a Java applet that you need only load into your Web browser for instant action.

Jalview is a Java applet and uses the same language as Dotlet (see Chapter 8) and has the same general properties:

- It runs on your own computer.
- When you load a sequence in Jalview, your sequence does not travel over the Internet; it stays in your computer.

If you want to ensure that none of your data travels across the Internet, choose the File➪Work Offline option on your browser as soon as Jalview is loaded.

Do not load confidential sequences in Jalview BEFORE doing this. The Web interface is NOT secure.

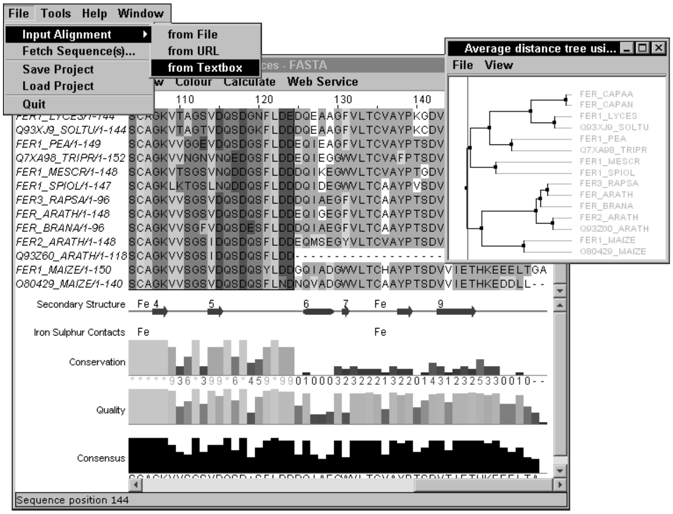
Starting Jalview

We only show you the basics about Jalview here. If you want to go further into discovering all the possibilities of this extremely powerful tool, we recommend you check out the extensive online documentation at `www.jalview.org/help.html`.

1. **Point your browser to** `www.jalview.org/download.html`**.**
2. **Click the Install button.**
3. **Accept every proposition prompted to you by the system.**

   If this is the first time you are using Jalview, it will need to upload the java applet. This will take a few minutes. It will then ask you to accept a few permissions having to do with network access — permissions you should graciously grant. If you have already used Jalview before, your computer will only check that you are using the latest version.
4. **Close ALL the windows that appear within the Jalview Window, as they only contain sample data.**
5. **Select the File➪Input Alignment➪From Textbox option, as shown in Figure 10-6.**

   If your alignment is already in a file, you can upload it with the From File option.
6. **Cut and paste your multiple alignment into the text box that appears (Figure 10-7) and click on the Accept box.**



```
                    110       120       130       140
FER1_LYCES/1-144    SCAGKVTAGSVDQSDGNFLDEDQEAAGFVLTCVAYPKGDV
Q93XJ9_SOLTU/1-144  SCAGKVTAGTVDQSDGKFLDDDQEAAGFVLTCVAYPKCDV
FER1_PEA/1-149      SCAGKVVGGEVDQSDGSFLDDEQIEAGFVLTCVAYPTSDV
Q7XA98_TRIPR/1-152  SCAGKVVNGNVNQEDGSFLDDEQIEGGWVLTCVAFPTSDV
FER1_MESCR/1-148    SCAGKVTSGSVNQDDGSFLDDDQIKEGWVLTCVAYPTGDV
FER1_SPIOL/1-147    SCAGKLKTGSLNQDDQSFLDDDQIDEGWVLTCAAYPVSDV
FER3_RAPSA/1-96     SCAGKVVSGSVDQSDQSFLDDDQIAEGFVLTCAAYPTSDV
FER_ARATH/1-148     SCAGKVVSGSVDQSDQSFLDDEQIGEGFVLTCAAYPTSDV
FER_BRANA/1-96      SCAGKVVSGFVDQSDESFLDDDQIAEGFVLTCAAYPTSDV
FER2_ARATH/1-148    SCAGKVVSGSIDQSDQSFLDDEQMSEGYVLTCVAYPTSDV
Q93Z60_ARATH/1-118  SCAGKVVSGSIDQSDQSFLDD-------------------
FER1_MAIZE/1-150    SCAGKVVSGSVDQSDQSYLDDGQIADGWVLTCHAYPTSDVVIETHKEEELTGA
O80429_MAIZE/1-140  SCAGKVVSGSVDQSDQSFLNDNQVADGWVLTCAAYPTSDVVIETHKEDDLL--
Conservation        *****936*399*6*459*9901000323222132201431232533001 0--
Sequence position 144
```

Figure 10-6: Uploading a multiple sequence alignment in Jalview.

You can either cut and paste your own multiple sequence alignment or you can pick up a dummy alignment at `www.tcoffee.org/dummy_aln.html`.

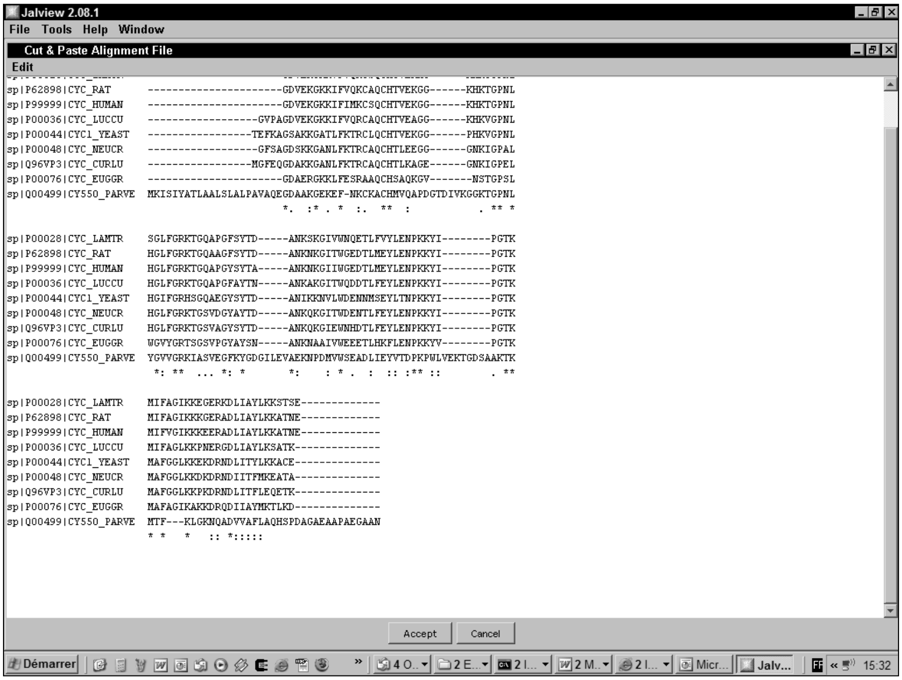
When you click on the accept box, a new window appears that contains your alignment, ready to be edited (Figure 10-8). This is the Cut & Paste Input window, and it contains the alignment you have entered in the box.

7. **Select the ClustalX Color Scheme in the Color pull-down Menu (Figure 10-8).**

   Choosing a color scheme is really a matter of taste; most multiple-sequence-alignment specialists can spend hours explaining to you why they prefer a special color scheme to every other alternative in the galaxy. Whatever your choice, any time you show a color alignment, you must be ready to answer the most common question in the world of biology:

   "What is the color scheme in this alignment?"

   To know more about the color schemes available in Jalview, read the excellent online documentation for this program at `www.jalview.org/help.html`; you'll find an extensive section on color schemes there. In Jalview, all the available color schemes are under the Colour menu (Figure 10-8). The ClustalX, initially designed by Tobby Gibson at the European Molecular Biology Laboratory, is probably one of the most popular.



```
sp|P62898|CYC_RAT       ----------------------GDVEKGKKIFVQKCAQCHTVEKGG------KHKTGPNL
sp|P99999|CYC_HUMAN     ----------------------GDVEKGKKIFIMKCSQCHTVEKGG------KHKTGPNL
sp|P00036|CYC_LUCCU     ------------------GVPAGDVEKGKKIFVQRCAQCHTVEAGG------KHKVGPNL
sp|P00044|CYC1_YEAST    -----------------TEFKAGSAKKGATLFKTRCLQCHTVEKGG------PHKVGPNL
sp|P00048|CYC_NEUCR     ------------------GFSAGDSKKGANLFKTRCAQCHTLEEGG------GNKIGPAL
sp|Q96VP3|CYC_CURLU     -----------------MGFEQGDAKKGANLFKTRCAQCHTLKAGE------GNKIGPEL
sp|P00076|CYC_EUGGR     ----------------------GDAERGKKLFESRAAQCHSAQKGV-------NSTGPSL
sp|Q00499|CY550_PARVE   MKISIYATLAALSLALPAVAQEGDAAKGEKEF-NKCKACHMVQAPDGTDIVKGGKTGPNL
                                              *.  :* . *  :.  **  :          . ** *

sp|P00028|CYC_LAMTR     SGLFGRKTGQAPGFSYTD-----ANKSKGIVWNQETLFVYLENPKKYI--------PGTK
sp|P62898|CYC_RAT       HGLFGRKTGQAAGFSYTD-----ANKNKGITWGEDTLMEYLENPKKYI--------PGTK
sp|P99999|CYC_HUMAN     HGLFGRKTGQAPGYSYTA-----ANKNKGIIWGEDTLMEYLENPKKYI--------PGTK
sp|P00036|CYC_LUCCU     HGLFGRKTGQAPGFAYTN-----ANKAKGITWQDDTLFEYLENPKKYI--------PGTK
sp|P00044|CYC1_YEAST    HGIFGRHSGQAEGYSYTD-----ANIKKNVLWDENNMSEYLTNPKKYI--------PGTK
sp|P00048|CYC_NEUCR     HGLFGRKTGSVDGYAYTD-----ANKQKGITWDENTLFEYLENPKKYI--------PGTK
sp|Q96VP3|CYC_CURLU     HGLFGRKTGSVAGYSYTD-----ANKQKGIEWNHDTLFEYLENPKKYI--------PGTK
sp|P00076|CYC_EUGGR     WGVYGRTSGSVPGYAYSN-----ANKNAAIVWEEETLHKFLENPKKYV--------PGTK
sp|Q00499|CY550_PARVE   YGVVGRKIASVEGFKYGDGILEVAEKNPDMVWSEADLIEYVTDPKPWLVEKTGDSAAKTK
                         *: **  ... *: *       *:    : * .  :  :: :** ::        . **

sp|P00028|CYC_LAMTR     MIFAGIKKEGERKDLIAYLKKSTSE-------------
sp|P62898|CYC_RAT       MIFAGIKKKGERADLIAYLKKATNE-------------
sp|P99999|CYC_HUMAN     MIFVGIKKKEERADLIAYLKKATNE-------------
sp|P00036|CYC_LUCCU     MIFAGLKKPNERGDLIAYLKSATK--------------
sp|P00044|CYC1_YEAST    MAFGGLKKEKDRNDLITYLKKACE--------------
sp|P00048|CYC_NEUCR     MAFGGLKKDKDRNDIITFMKEATA--------------
sp|Q96VP3|CYC_CURLU     MAFGGLKKPKDRNDLITFLEQETK--------------
sp|P00076|CYC_EUGGR     MAFAGIKAKKDRQDIIAYMKTLKD--------------
sp|Q00499|CY550_PARVE   MTF---KLGKNQADVVAFLAQHSPDAGAEAAPAEGAAN
                        * *   *   :: *:::::
```

Figure 10-7: The Jalview Alignment TextBox.

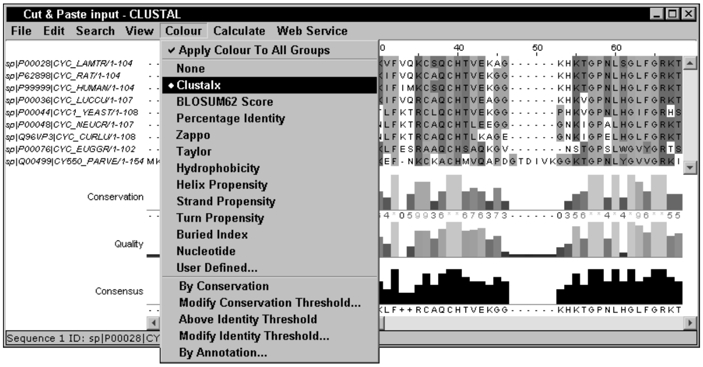

Cut & Paste input - CLUSTAL
File Edit Search View Colour Calculate Web Service
✓ Apply Colour To All Groups
None
◆ Clustalx
BLOSUM62 Score
Percentage Identity
Zappo
Taylor
Hydrophobicity
Helix Propensity
Strand Propensity
Turn Propensity
Buried Index
Nucleotide
User Defined...
By Conservation
Modify Conservation Threshold...
Above Identity Threshold
Modify Identity Threshold...
By Annotation...
sp|P00028|CYC_LAMTR/1-104
sp|P62898|CYC_RAT/1-104
sp|P99999|CYC_HUMAN/1-104
sp|P00036|CYC_LUCCU/1-107
sp|P00044|CYC1_YEAST/1-108
sp|P00048|CYC_NEUCR/1-107
sp|Q96VP3|CYC_CURLU/1-108
sp|P00076|CYC_EUGGR/1-102
sp|Q00499|CY550_PARVE/1-154
Conservation
Quality
Consensus
Sequence 1 ID: sp|P00028|CY

Figure 10-8: Changing the Jalview Color Scheme.

Editing a group of sequences

When you edit an alignment, you usually want to keep some of your sequences aligned the way they are relative to one another. What you want to do is *collectively* modify their alignment. To do this, you need to define them as a group, as follows:

1. **For the purposes of this example, input the dummy alignment we describe in the previous section into Jalview.**

   You can grab the dummy alignment from `www.tcoffee.org/dummy_aln.html`.

2. **Keep the Ctrl key pressed while you click sequences 1, 2, and 4 to select them, as shown in Figure 10-9.**

   Selecting the sequences on a phylogenetic tree is also possible:

   a. Select Calculate➪Calculate Tree ➪Neighbor Joining Tree Using PID.

   b. On the tree that pops up, you can select sequences individually (Figure 10-10).

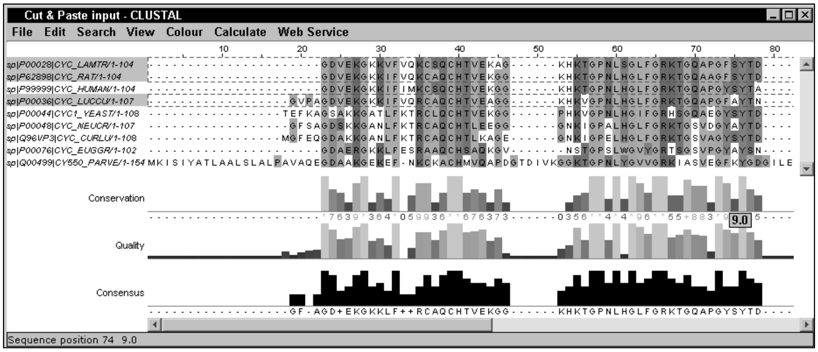
3. **To insert/remove gaps:**

   a. Keep the Ctrl key pressed.

   b. Put your mouse pointer right where you want to insert or remove the gap.

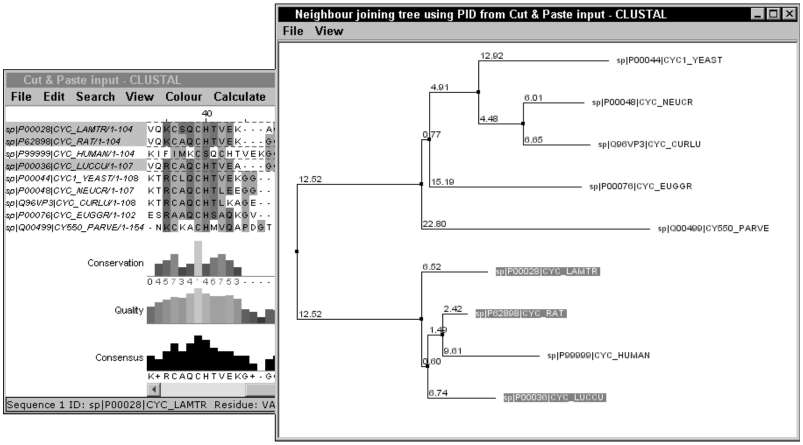
   c. Drag to the left or to the right to shift your sequences, as shown in Figure 10-11.

   You can edit one sequence at a time by pressing the Shift key instead of Ctrl.


Cut & Paste input - CLUSTAL
File Edit Search View Colour Calculate Web Service
sp|P00028|CYC_LAMTR/1-104
sp|P62898|CYC_RAT/1-104
sp|P99999|CYC_HUMAN/1-104
sp|P00036|CYC_LUCCU/1-107
sp|P00044|CYC1_YEAST/1-108
sp|P00048|CYC_NEUCR/1-107
sp|Q96VP3|CYC_CURLU/1-108
sp|P00076|CYC_EUGGR/1-102
sp|Q00499|CY550_PARVE/1-154
Conservation
Quality
Consensus
9.0

Sequence position 74 9.0

Figure 10-9: Selecting sequences in Jalview.



Sequence 1 ID: sp|P00028|CYC_LAMTR Residue: VA

Figure 10-10: Selecting sequences on a phylogenetic tree.

4. **To remove empty columns, use Edit➪Remove empty columns.**

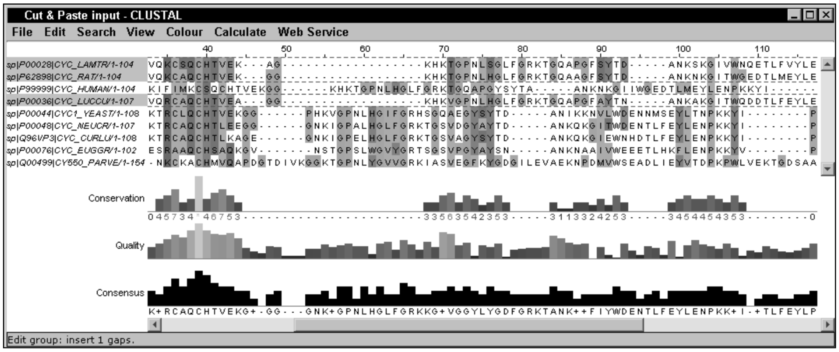
   Empty columns (containing only gaps) may appear while you edit. This simple command gets rid of them for you.

5. **Choose Edit➪Pad Gaps to ensure that no empty column appears at the end of the multiple sequence alignment.**

   When you shift some sequences, you insert gaps. If you don't then edit the gaps at the end of the alignment, you will have problems using your alignment in other applications, as your sequences will seem to be unaligned.

6. **To edit another alignment, go back to the main Jalview window.**

   The File menu in the main window and the File menu in the alignment window do not display the same functionalities.



```
sp|P00028|CYC_LAMTR/1-104     VQKCSQCHTVEK---AG------------------KHKTGPNLSGLFGRKTGQAPGFSYTD-----ANKSKGIVWNQETLFVYLE
sp|P62898|CYC_RAT/1-104       VQKCAQCHTVEK---GG------------------KHKTGPNLHGLFGRKTGQAAGFSYTD-----ANKNKGITWGEDTLMEYLE
sp|P99999|CYC_HUMAN/1-104     KIFIMKCSQCHTVEKGG------KHKTGPNLHGLFGRKTGQAPGYSYTA-----ANKNKGIIWGEDTLMEYLENPKKYI------
sp|P00036|CYC_LUCCU/1-107     VQRCAQCHTVEA---GG------------------KHKVGPNLHGLFGRKTGQAPGFAYTN-----ANKAKGITWQDDTLFEYLE
sp|P00044|CYC1_YEAST/1-108    KTRCLQCHTVEKGG------PHKVGPNLHGIFGRHSGQAEGYSYTD-----ANIKKNVLWDENNMSEYLTNPKKYI--------P
sp|P00048|CYC_NEUCR/1-107     KTRCAQCHTLEEGG------GNKIGPALHGLFGRKTGSVDGYAYTD-----ANKQKGITWDENTLFEYLENPKKYI--------P
sp|Q96VP3|CYC_CURLU/1-108     KTRCAQCHTLKAGE------GNKIGPELHGLFGRKTGSVAGYSYTD-----ANKQKGIEWNHDTLFEYLENPKKYI--------P
sp|P00076|CYC_EUGGR/1-102     ESRAAQCHSAQKGV-------NSTGPSLWGVYGRTSGSVPGYAYSN-----ANKNAAIVWEEETLHKFLENPKKYV--------P
sp|Q00499|CY550_PARVE/1-154   -NKCKACHMVQAPDGTDIVKGGKTGPNLYGVVGRKIASVEGFKYGDGILEVAEKNPDMVWSEADLIEYVTDPKPWLVEKTGDSAA

                              045734*46753-----------------------33563542353-----3113324253-----3454454353--------0

                              K+RCAQCHTVEKG+-GG---GNK+GPNLHGLFGRKKG+VGGYLYGDFGRKTANK++FIYWDENTLFEYLENPKK+I-+TLFEYLP

Edit group: insert 1 gaps.
```

Figure 10-11: Effect of dragging a group to the right.

Useful features of Jalview

Although explaining all the possibilities of Jalview is clearly beyond the scope of this chapter, we encourage you to explore them. Jalview is like all the programs we introduce here: You can't break anything, so experiment away!

WARNING!

That said, remember that it is easy to ruin a good alignment. Save intermediate results and do not hesitate to use the magic undo button (Ctrl+Z).

Table 10-4 lists some of the features we find especially useful.

Table 10-4 Some Useful Features of Jalview

Command	*Description*
Calculate➪ Autocalculate Consensus	Automatic update of the graph below the alignment. This graph — looking like a city skyline — indicates the level of conservation within the alignment. If you set this option, the graph is updated automatically while you're editing.
Edit➪Remove Redundancy	Makes sure that no pair of sequences is more than *x* percent similar.
Calculate➪Tree➪ Neighbor Joining Tree Using PID	Computes and displays a phylogenetic tree in graphic format on which you can select sequences for group editing.

Saving your alignment in Jalview

The new Jalview makes it very easy to save either a colored version of your alignment or a simple text version. We recommend you save both. The colored version is useful for making illustrations while the text version is necessary if you want to keep working with your alignment.

To save your multiple alignment in text and HTML (for all the pretty colors) format, follow these steps:

1. **From the Jalview Alignment Window, choose File➪Save Alignment AS.**
2. **In the dialog box that appears, select FASTA in the File Type pull-down menu.**

   We recommend you either select the FASTA format (most portable) or the CLUSTAL format (ALN), which is more easily read by us humans.
3. **From the Jalview Alignment Window, choose File➪Export➪HTML.**

   Doing so saves a colorized version of your alignment that you can visualize with any Web browser (Mozilla, Netscape).

Preparing Your Multiple Alignment for Publication

Showtime has finally come: You have the multiple alignment you want, and you're determined to show the world! You want to show this alignment to your colleagues, you want to include it in publications, on posters, in mails, and maybe on T-shirts, mugs, and other merchandise — you're even mulling over the idea of a tattoo. In short, you need a high-impact picture to convince people that your research is going gangbusters. In this section, we list a few online utilities that you can use to beautify your multiple sequence alignments and make them look sharp as a tack.

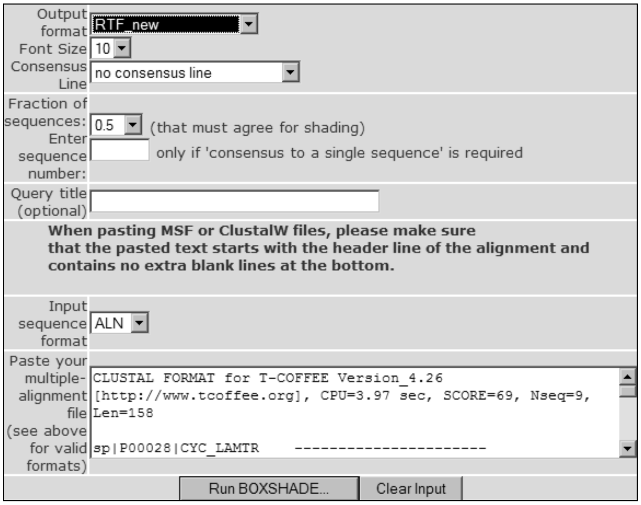
Using Boxshade

Boxshade is a utility that allows you to put some life into your alignment. It shades columns according to their level of conservation and produces files that you can easily manipulate for inclusion in reports or articles. (We give you other powerful tools of the same kind in Table 10-7.)

To get the ball rolling with Boxshade, do the following:

1. **Point your browser to** `www.ch.embnet.org/software/BOX_form.html`.

   The Boxshade page of `ch.EMBnet.org` appears, part of which you can see in Figure 10-12.



Output format: RTF_new

Font Size: 10

Consensus Line: no consensus line

Fraction of sequences: 0.5 (that must agree for shading)

Enter sequence number: only if 'consensus to a single sequence' is required

Query title (optional):

When pasting MSF or ClustalW files, please make sure that the pasted text starts with the header line of the alignment and contains no extra blank lines at the bottom.

Input sequence format: ALN

Paste your multiple-alignment file (see above for valid formats):

```
CLUSTAL FORMAT for T-COFFEE Version_4.26
[http://www.tcoffee.org], CPU=3.97 sec, SCORE=69, Nseq=9,
Len=158

sp|P00028|CYC_LAMTR    ----------------------
```

Figure 10-12: Part of the Boxshade home page.

2. **Choose RTF_new from the Output Format drop-down menu.**

   Most of these formats are difficult to use on a PC. The most convenient is RTF New (Rich Text Format). It generates a file that most word processors (such as Microsoft Word) can read.

3. **Select the font size you want.**

   If your alignment is long, select a small font size. If you have selected RTF_new as your output, you can change your font size later.

4. **Choose Add a Consensus Line with Letters from the Consensus Line drop-down menu.**

   This adds a consensus sequence that contains the most common amino acid for each column.

5. **Select the fraction of sequences you would like shaded.**

   Boxshade only shades columns that have a certain level of conservation. This feature ensures that conserved columns show up in your alignment. If you request 0.5, for example, it means you want half the amino acids (or nucleotides) to be conserved for some shading to occur.

   Conservation doesn't necessarily mean *identity* in Boxshade. Similar residues, such as isoleucine and valine, also account for conservation. Two types of shading exist:

   - *Black:* Identical amino acids or nucleotides
   - *Gray:* Similar amino acids

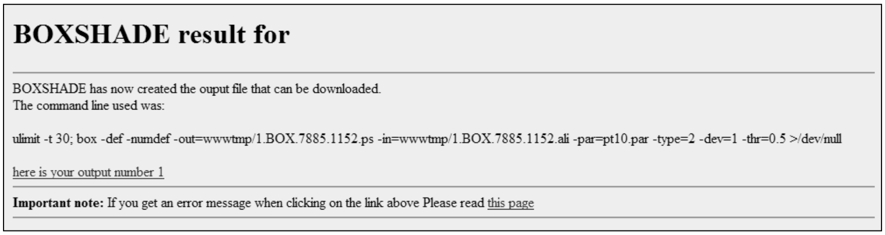
6. **Select the format of the multiple sequence alignment you want to use.**

   If you're planning to use the dummy multiple alignment from `www.tcoffee.org/dummy_aln.html`, select ALN.

   Other possible formats include anything that the READSEQ program can read, such as FASTA or PIR alignments.

7. **Paste your multiple sequence alignment into the Sequence window.**
8. **Click the Run Boxshade button.**

   An intermediate page appears, as shown in Figure 10-13.



BOXSHADE result for

BOXSHADE has now created the ouput file that can be downloaded.
The command line used was:

```
ulimit -t 30; box -def -numdef -out=wwwtmp/1.BOX.7885.1152.ps -in=wwwtmp/1.BOX.7885.1152.ali -par=pt10.par -type=2 -dev=1 -thr=0.5 >/dev/null
```

here is your output number 1

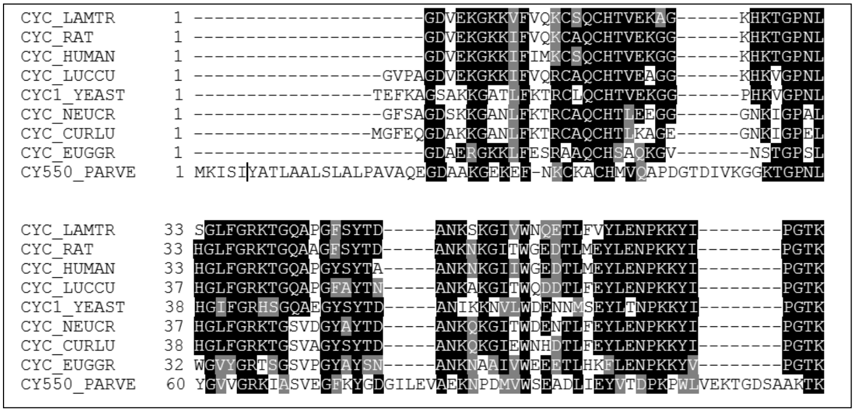
Important note: If you get an error message when clicking on the link above Please read this page

Figure 10-13: The Boxshade intermediate page.

If this page does not contain an active hyperlink, it means that Boxshade had troubles reading your alignment. (Refer to Figure 10-13 for an example of an active hyperlink.) When this happens, double-check the alignment format. If this doesn't help, use a different format.

9. **Click the here is your output link, save to a local file, then open the local file with Word.**

   Figure 10-14 shows you what your multiple sequence alignment looks like after proper shading.



```
CYC_LAMTR      1 ----------------------GDVEKGKKVFVQKCSQCHTVEKAG------KHKTGPNL
CYC_RAT        1 ----------------------GDVEKGKKIFVQKCAQCHTVEKGG------KHKTGPNL
CYC_HUMAN      1 ----------------------GDVEKGKKIFIMKCSQCHTVEKGG------KHKTGPNL
CYC_LUCCU      1 ------------------GVPAGDVEKGKKIFVQRCAQCHTVEAGG------KHKVGPNL
CYC1_YEAST     1 -----------------TEFKAGSAKKGATLFKTRCLQCHTVEKGG------PHKVGPNL
CYC_NEUCR      1 ------------------GFSAGDSKKGANLFKTRCAQCHTLEEGG------GNKIGPAL
CYC_CURLU      1 -----------------MGFEQGDAKKGANLFKTRCAQCHTLKAGE------GNKIGPEL
CYC_EUGGR      1 ----------------------GDAERGKKLFESRAAQCHSAQKGV-------NSTGPSL
CY550_PARVE    1 MKISIYATLAALSLALPAVAQEGDAAKGEKEF-NKCKACHMVQAPDGTDIVKGGKTGPNL


CYC_LAMTR     33 SGLFGRKTGQAPGFSYTD-----ANKSKGIVWNQETLFVYLENPKKYI--------PGTK
CYC_RAT       33 HGLFGRKTGQAAGFSYTD-----ANKNKGITWGEDTLMEYLENPKKYI--------PGTK
CYC_HUMAN     33 HGLFGRKTGQAPGYSYTA-----ANKNKGIIWGEDTLMEYLENPKKYI--------PGTK
CYC_LUCCU     37 HGLFGRKTGQAPGFAYTN-----ANKAKGITWQDDTLFEYLENPKKYI--------PGTK
CYC1_YEAST    38 HGIFGRHSGQAEGYSYTD-----ANIKKNVLWDENNMSEYLTNPKKYI--------PGTK
CYC_NEUCR     37 HGLFGRKTGSVDGYAYTD-----ANKQKGITWDENTLFEYLENPKKYI--------PGTK
CYC_CURLU     38 HGLFGRKTGSVAGYSYTD-----ANKQKGIEWNHDTLFEYLENPKKYI--------PGTK
CYC_EUGGR     32 WGVYGRTSGSVPGYAYSN-----ANKNAAIVWEEETLHKFLENPKKYV--------PGTK
CY550_PARVE   60 YGVVGRKIASVEGFKYGDGILEVAEKNPDMVWSEADLIEYVTDPKPWLVEKTGDSAAKTK
```

Figure 10-14: Boxshade output.

Logos

Logos are a terrific way to generate high-impact pictures from multiple sequence alignments. They have been around in sequence analysis for quite some time. Figure 10-15 shows a logo representation of the alignment in Figure 10-1. Notice how the conserved amino acids (such as cysteines) stick out, indicating regions of potential biological importance.

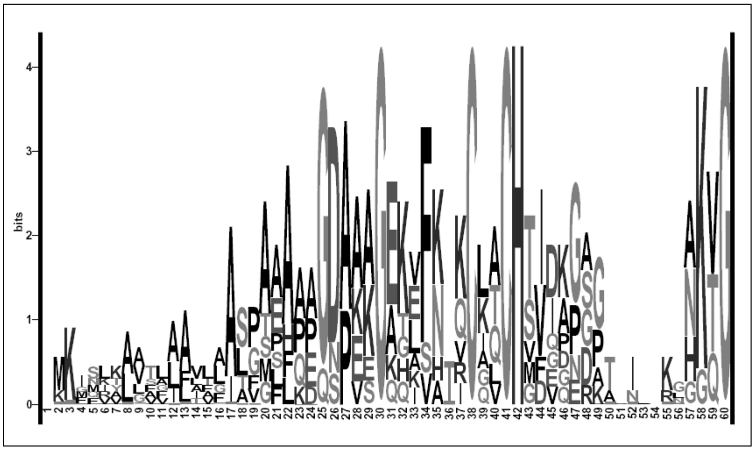


Figure 10-15: Logo representation of a multiple sequence alignment.

When looking at a sequence logo, you can consider the following elements:

- Each position corresponds to a column in the multiple alignment.
- The total height of a logo position depends on the degree of conservation in the corresponding multiple alignment column. Very conserved alignment columns give you high logo positions. Positions that contain a very heterogeneous mixture of symbols yield low logo positions.
- The size of each letter in a logo position depends on how frequent this letter is in the column.
- The top letter is always the most frequent in the column.

It's possible to build logos with either protein or nucleotide multiple sequence alignments. Of course, the better the alignment, the more informative the logo. The most complete logo server around is the one from Berkeley (at `weblogo.berkeley.edu/`) — and it is probably the easiest-to-use server available today. You simply cut and paste your alignment in any convenient format (such as ALN or FASTA) and the server returns a logo like the one in Figure 10-15. A slightly more complicated alternative is also available from: `www.cbs.dtu.dk/~gorodkin/appl/plogo.html`.

Logos make sense only if you have a nice block with a few highly conserved positions surrounded by highly degenerated positions. There is a handy utility on the Web that identifies blocks within your multiple alignments and turns each of them into a log. Check it out at

```
blocks.fhcrc.org./blocks/process_blocks.html
```

Editing and Analyzing Multiple Sequence Alignments for Free over the Internet

This last section briefly introduces and/or recapitulates some lists of resources available on the Internet for editing, analyzing, or visualizing your multiple sequence alignments. If you need to reformat your multiple sequence alignment, you can use the resources indicated in Table 10-2.

Finding multiple-sequence-alignment editors

The only tool available for editing multiple sequence alignments online is Jalview. If you're ready to install some program locally on your machine, a vast selection of high-quality packages exists. Table 10-5 lists some of these programs; most come with extensive documentation for installation and usage.

Table 10-5 Packages for Editing Multiple Sequence Alignments

Name	*Address*	*Description*
Jalview	`www.jalview.org` `www.es.embnet.org/Services/MolBio/jalview/`	Java package, available online
Kalignview	`msa.cgb.ki.se`	Nice online alignment viewer
CINEMA	`www.bioinf.man.ac.uk/dbbrowser/CINEMA2.1/`	A very complete Java package
Seaview	`pbil.univ-lyon1.fr/software/seaview.html`	A beautiful editor, very easy to install
Belvu	`www.cgr.ki.se/cgr/groups/sonnhammer/Belvu.html`	Useful for removing redundancy

(continued)

Table 10-5 *(continued)*

Name	*Address*	*Description*
Bioedit	`www.mbio.ncsu.edu/BioEdit/bioedit.html`	Adapted for RNA
RALEE	`www.sanger.ac.uk/Users/sgj/ralee/`	An RNA viewer
Review	`bioweb.pasteur.fr/cgi-bin/seqanal/review-edital.pl`	A very complete list of viewers

Finding tools to interpret your multiple sequence alignment

The interpretation of a multiple alignment depends very much on its appearance. Some tools on the Net can help you make sense of your multiple alignments by extracting blocks or singling out special positions. Table 10-6 lists some of these resources.

Table 10-6 Extracting Information from a Multiple Sequence Alignment

Name	*Address*	*Description*
Logo	`weblogo.berkeley.edu`, `www-lmmb.ncifcrf.gov/~toms/sequencelogo.html`, `www.cbs.dtu.dk/~gorodkin/appl/plogo.html`	Logos
Blocks	`blocks.fhcrc.org/blocks/process_blocks.html`	Identifies blocks
Blockgap	`www.bork.embl-heidelberg.de/Alignment/blockgap.html`	Measures blocks
Lama	`blocks.fhcrc.org/blocks-bin/LAMA_search.sh`	Compares your multiple alignment with the BLOCKs database
Amas	`www.compbio.dundee.ac.uk/servers/amas_server.html`	Identifies important features in the multiple alignment

Finding tools for beautifying your multiple alignments

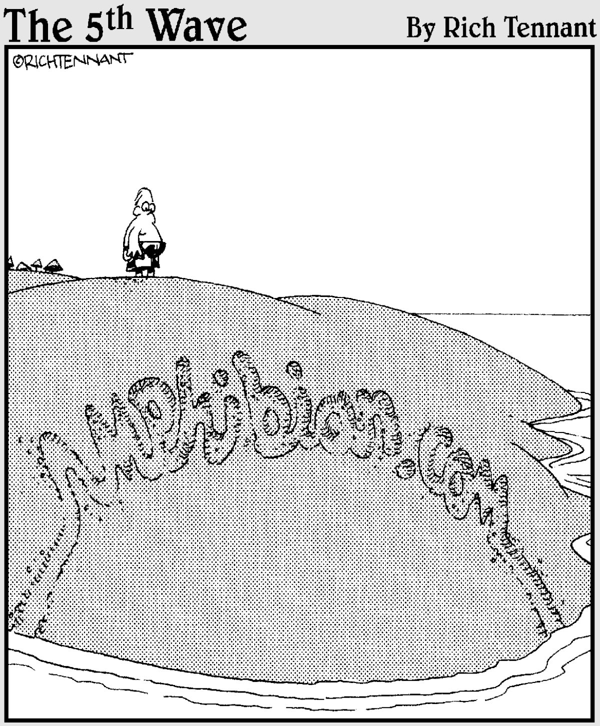
Boxshade — which we describe earlier in this chapter — is one of the simplest tools for improving the visual aspect of a multiple sequence alignment. It isn't the only server you can use for this purpose, however. Table 10-7 lists some alternative resources that are available online.

Table 10-7 Multiple Alignment Beautifying Tools

Name	*Address*	*Description*
ESPript	`espript.ibcp.fr`	A very powerful shading-and-coloring tool
Boxshade	`www.ch.embnet.org/software/BOX_form.html`	Shading in black and white
Mview	`bioweb.pasteur.fr/seqanal/interfaces/mview_blast-simple.html`	Can process BLAST alignments

Part IV

Becoming a Specialist: Advanced Bioinformatics Techniques



In this part . . .

This part is dedicated to a major slice of bioinformatics: predictions. Accurate predictions are the ultimate dream of biologists. It is a very practical dream because computers are cheap whereas experiments are expensive.

In this part we introduce a Biology Holy Grail: the prediction of protein structures. We also present you with its more successful nucleic acid counterpart: the prediction of RNA structure. We call this sort of stuff *advanced* bioinformatics because the results of these methods only make sense if you know how to interpret them and if you have a good understanding of their limitations. These aren't methods that you can trust blindly, and in this part, we tell you when to keep your eyes peeled!

Chapter 11

Working with Protein 3-D Structures

In This Chapter

- Looking at 3-D structures and why they're important
- Predicting the secondary structure of a protein
- Finding the best 3-D model for your sequence
- Creating pictures of 3-D protein structures for your talks and reports
- Analyzing a sequence from a structural point of view
- Displaying interactive (rotating) 3-D molecular models on your own PC

> *There is no foreseeable limit to the number of proteins whose structure is worth analyzing, since each will have its own unique function which demands explanation in structural terms.*
>
> — John C. Kendrew, Nobel Lecture 1962

When you study a protein, you are usually interested in its function. One thing all biologists know these days is that there is a tight relationship between the structure and the function of a protein. As soon as you find an interesting feature at the level of the protein sequence — perhaps a motif or an evolutionarily conserved segment — the right thing to do is to ask what this feature looks like within the 3-D structure.

Typical questions you want to ask are

- Do the amino acids at given sequence positions contribute to the stability of the protein structure?
- Why is this sequence segment conserved (or variable)?

- Are these amino acids located at the surface of my protein molecules?
- Are these amino acids directly involved in the function of the protein?
- Are these amino acids involved in the binding of another molecule?

In this chapter, we show you how to look at your sequence from a 3-D perspective.

From Primary to Secondary Structures

Biologists refer to the sequence of a protein as its *primary structure,* as opposed to the 3-D or *tertiary* structure that is the final shape of the protein. No prize for guessing that an intermediary level of structure exists that is known as the *secondary* structure!

When the first crystallographers started looking at protein structures, they discovered (and predicted) that there was a hierarchy in the way that amino acid sequences fold onto themselves to become a biologically active molecule. Amino acids first look to their immediate neighbors in the sequence to form regions of regular, periodical conformations (backbone shapes). After this is done, the chain collapses further by folding those preshaped regions onto each other (or onto unstructured regions), leading to the final 3-D structure in which residues that are far apart in the sequence come in direct contact with each other.

There are three types of local segments:

- **Helices:** Where residues seem to be following the shape of a spring. The most common are the so-called alpha helices.
- **Extended or Beta-strands:** Where residues are in line and successive residues turn their back to each other.
- **Random coils:** When the amino-acid chain is neither helical nor extended.

Within the latter category, biologists like to distinguish those cases where the chain makes a sharp turn (90° or more) by labeling them *loops.*

Predicting the secondary structure of a protein sequence

Predicting the secondary structure of proteins was one of the hottest goals of the 1990s. It is fair to say that this is one of the great successes of that decade. Nowadays, fairly good servers are available that use Hidden Markov Models and neural networks to accurately predict the secondary structure of any protein that may interest you. In this section, we show you how to use PSIPRED, one of the most accurate of these servers.

If your protein has enough homologues in the current databases, you can expect its secondary structure prediction to be close to 80 percent accurate. However, bear in mind that this is *only* a prediction; as with all predictions, it can be more or less inaccurate.

To get started using PSIPRED, do the following:

1. **Prepare your favorite protein in a FASTA-formatted .txt file.**

   We used GenPep NP_360043, *Rickettsia conorii* sequence TolB for this example. You can fetch it from the NCBI Web server (`www.ncbi.nlm.nih.gov`); see Step 1 in the later section "Looking at sequence features in 3-D."

2. **Point your browser to** `bioinf.cs.ucl.ac.uk/psipred/`**.**

   This displays the home page of the Protein Structure Prediction server, which is maintained by the Bioinformatics Unit of University College in London (U.K.).

3. **Scroll down the page and click CLICK HERE TO ACCESS THE SERVER.**

   A one-page input form appears.

4. **Copy your sequence and paste it into the Input Sequence window.**

   Both FASTA and plain (with no header) formats are accepted here.

5. **Keep the default pre-selected option in the Choose Prediction Method section.**

6. **Keep the default pre-selected option in the Filtering Options.**

7. **Enter your e-mail address in the Submit Sequence section.**

   You'll also find a field in this section where you can enter a short name for your sequence; this is handy if you need help remembering what the analysis was all about.

8. **Click the Predict button.**

   This server returns its results by e-mail, usually in less than 30 minutes (sometimes more, depending on its load). By default, the server runs PSIPRED, the secondary structure prediction method.

   Alternatively, you can select three other prediction methods: one for transmembrane segments and two for fold recognition. Feel free to explore these if you have time.

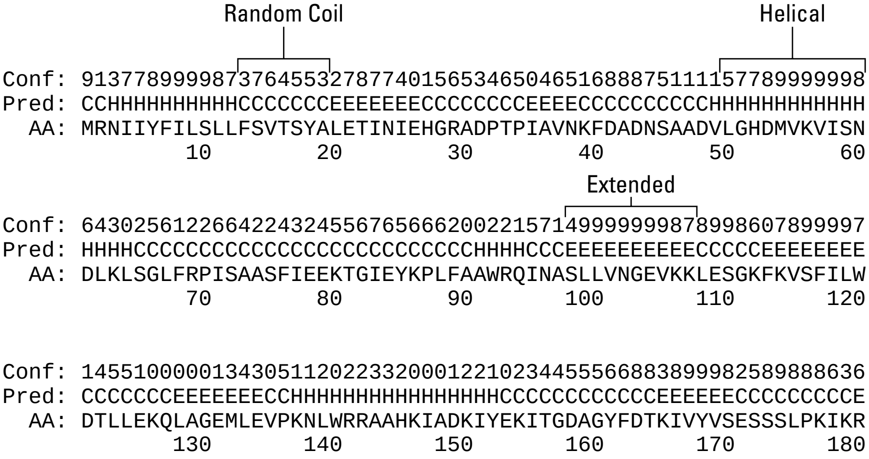
If everything is fine, a new page pops up and informs you that your prediction job has been submitted and that the result will be sent to the e-mail address you provided in Step 7 (check to make sure that the e-mail address is correct!).

It also tells you how many jobs are currently in front of you in the processing queue. At this point, there's little that you can do but wait for your e-mail to show up (about 20 minutes or more, depending on the server load). Go and have a coffee!

The server does not accept free, Web-based e-mail addresses (such as those from MSN's Hotmail or Google's Gmail). No results will be returned.

When you come back, look for an e-mail from `psipred@cs.ucl.ac.uk`, similar to what you see in Figure 11-1. The output is a simple text file in which each line of the sequence (AA) is now printed side by side with two other lines:

- The *prediction* line (Pred), consisting of H, E, and C characters, denoting the predicted conformation for each residue (H = *Helical,* E = *Extended,* and C = *Random Coil*).
- The *confidence* (Conf) line, consisting of digits from 9 to 0, indicating the reliability of the prediction for each position (9 = high, 0 = poor).



```
                   Random Coil                                Helical

Conf: 913778999987376455327877401565346504651688875111157789999998
Pred: CCHHHHHHHHHHCCCCCCCEEEEEEECCCCCCCCEEEECCCCCCCCCCHHHHHHHHHHHH
  AA: MRNIIYFILSLLFSVTSYALETINIEHGRADPTPIAVNKFDADNSAADVLGHDMVKVISN
              10        20        30        40        50        60
                                                   Extended

Conf: 643025612266422432455676566620022157149999999878998607899997
Pred: HHHHCCCCCCCCCCCCCCCCCCCCCCCCCCHHHHCCCEEEEEEEEEECCCCCEEEEEEEE
  AA: DLKLSGLFRPISAASFIEEKTGIEYKPLFAAWRQINASLLVNGEVKKLESGKFKVSFILW
              70        80        90       100       110       120


Conf: 145510000013430511202233200012210234455566883899982589888636
Pred: CCCCCCCEEEEEEECCHHHHHHHHHHHHHHHHCCCCCCCCCCCCEEEEEECCCCCCCCCE
  AA: DTLLEKQLAGEMLEVPKNLWRRAAHKIADKIYEKITGDAGYFDTKIVYVSESSSLPKIKR
             130       140       150       160       170       180
```

Figure 11-1: Typical PSIPRED e-mail: Partial output for R. conorii TolB (NP_360043).

To get the various columns well lined up as in Figure 11-1, copy the e-mailed results into a Word document and use a constant spaced font such as Courier 10.

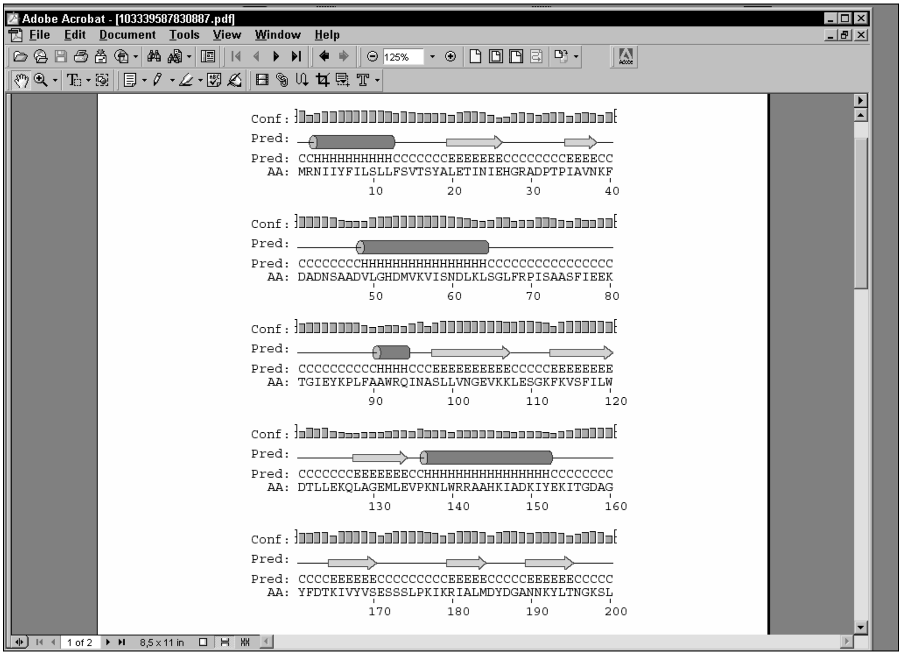
PSIPRED is very popular because of its accuracy (over 80 percent correctly predicted positions, on average) and also because it produces a nice graphic output that comes in handy for publication. To retrieve the graphic files, look at the very end of your PSIPRED e-mail. It reads like this:

```
Calculate PostScript, PDF and JPEG graphical output for
          this result using
http://bioinf2.cs.ucl.ac.uk/cgi-
          bin/psipred/gra/nphview2.cgi?id=xxxxx.psi
```

Clicking this URL gets you to a PSIPRED output page that offers you the opportunity to download the graphic files in a number of different formats:

- **PostScript File:** `http://bioinf2.cs.ucl.ac.uk/psiout/xxxxxx.ps`
- **PDF File:** `http://bioinf2.cs.ucl.ac.uk/psiout/xxxxxx.pdf`
- **JPEG Page 1:** `http://bioinf2.cs.ucl.ac.uk/psiout/xxxxx_1.jpg`
- **JPEG Page 2:** `http://bioinf2.cs.ucl.ac.uk/psiout/xxxxx_2.jpg`

If your computer is already set up with Adobe Acrobat, select the graphic file in PDF format by simply clicking the appropriate URL. That should open up a display similar to what's shown in Figure 11-2. You may also download the JPEG version if you want to incorporate it into a PowerPoint presentation.


Adobe Acrobat - [103339587830887.pdf]
File Edit Document Tools View Window Help
125%
1 of 2
8,5 x 11 in

```
Conf:
Pred:
Pred: CCHHHHHHHHHHCCCCCCEEEEEEEECCCCCCCEEEEECC
  AA: MRNIIYFILSLLFSVTSYALETINIEHGRADPTPIAVNKF
                10        20        30        40

Conf:
Pred:
Pred: CCCCCCCCHHHHHHHHHHHHHHHHCCCCCCCCCCCCCCCC
  AA: DADNSAADVLGHDMVKVISNDLKLSGLFRPISAASFIEEK
                50        60        70        80

Conf:
Pred:
Pred: CCCCCCCCCCHHHHCCCEEEEEEEEEECCCCCEEEEEEEE
  AA: TGIEYKPLFAAWRQINASLLVNGEVKKLESGKFKVSFILW
                90        100       110       120

Conf:
Pred:
Pred: CCCCCCCEEEEEEECCHHHHHHHHHHHHHHHHCCCCCCCC
  AA: DTLLEKQLAGEMLEVPKNLWRRAAHKIADKIYEKITGDAG
                130       140       150       160

Conf:
Pred:
Pred: CCCCEEEEEECCCCCCCCCEEEEECCCCCEEEEEECCCCC
  AA: YFDTKIVYVSESSSLPKIKRIALMDYDGANNKYLTNGKSL
                170       180       190       200
```

Figure 11-2: Typical PSIPRED graphics: Partial output for R. conorii TolB (NP_360043).

The graphic output is organized along the same lines as the text output, with the confidence line replaced by a confidence *profile,* and the Prediction line completed by explicit symbols (Helical: *cylinder,* Extended: *arrow,* Coil: *nothing*).

Predicting additional structural features

PSIPRED is ideal if you need a quick answer to a basic question, such as *What is the secondary structure of my protein?* If you want to know more, you must turn to other sites capable of predicting a much broader variety of features in your protein.

The PredictProtein server is probably the most comprehensive site for protein structure analysis. Unfortunately, the original site (based at Columbia University) is also very busy — and response times can be more than a day or two. Luckily, the PredictProtein server has numerous mirror sites throughout the world — including Europe, the United States, Asia, and Australia (but beware: some of the URLs listed on the server's home page are obsolete). Here is a list of up-to-date URLs:

Europe:	`www.predictprotein.org/`
	`www.cmbi.ru.nl/bioinf/predictprotein/`
USA:	`cubic.bioc.columbia.edu/predictprotein/`
	`www.sdsc.edu/predictprotein/`
Asia:	`www.cbi.pku.edu.cn/predictprotein/`

Your basic PredictProtein server

To check out a PredictProtein server, have your protein sequence handy in another browser window or in a local file, and follow these steps:

1. **Point your browser to** `www.sdsc.edu/predictprotein/`**.**

   The PredictProtein Server page dutifully appears.

   We found (for now!) that the San Diego site tends to be the more responsive.

2. **Click the** Submit a Protein Sequence for Prediction **link (the first bullet).**

   A default submission form appears.

3. **Enter your e-mail address in the top text field.**

4. **Select the output format by checking a box.**

   In the default mode (Results on Our Web Site), the server sends you an e-mail that indicates the URL of your result file on the PredictProtein site. This file remains available for *one day.* It can be in HTML or HTML for Printouts. In this mode, a graphic processing of the results is also offered.

If you deselect the default mode, the results are sent to you as a rather long e-mail. Beware that your local e-mail device might not be able to handle these large result files.

So, in all cases, the server sends you an e-mail; with an address where to find your results or with the results themselves.

5. **Copy your sequence from another browser window (or your local file) and paste it into the Paste or Type Your Sequence box.**

   The server only accepts a single protein sequence, in one-letter code, in the raw (plain) format. No FASTA header.

6. **If you like, you can enter a short name for your sequence in the One-Line Name of Protein text field.**

7. **Click the Submit/Run Prediction button.**

Now all you have to do is check your e-mail from time to time. To estimate how long that will take, click the yellow Wait button at the bottom-left of the form to display the status of the PredictProtein queue.

A default analysis returns the following:

- A secondary structure prediction on the three conformational states (H = *Helical,* E = *Extended,* and C = *Random Coil*)
- A prediction of the solvent accessibility of the various residues
- A prediction of transmembrane helices and their topologies
- A prediction of globular regions in your protein
- A prediction of the coiled/coil regions of your protein
- A description of the PROSITE motifs matching your sequence (for more on PROSITE motifs, see Chapter 6)
- A description of the putative domain structure for your protein (Prodom domains)
- A list of likely homologues of your query sequence found in the Swiss-Prot protein database using the similarity-search program PSI-BLAST
- A multiple alignment of your query with these homologous sequences generated by the MaxHom program
- A prediction of bound cysteines (disulphide bonds) in your sequence
- A description of the composition-biased regions in your sequence

These various features are reported only if they occur in your query sequence. If you need the details on any of these predictions, you can run additional programs by using the Advanced or Expert input form.

The PredictProtein Meta server

The PredictProtein site also runs a META server — an interface that lets you submit your sequence to many servers at once, and which then automatically retrieves the results for you. Here's how it's done:

1. **Point your browser to** `www.sdsc.edu/predictprotein/`.

   You're back at the UC-San Diego PredictProtein Server page.

2. **Click the** Submit Requests to META-PP **link (the second bullet).**

   This displays a one-page input form.

3. **Enter your e-mail address in the top text field.**

   This server returns its results by e-mail.

4. **Copy your sequence from another browser window (or from a local file) and paste it into the Paste or Type Your Sequence box.**

   A one-letter code sequence given in the raw (plain) format is fine.

5. **Scroll down to find the table of options available and choose the methods that you're interested in.**

   Most of these services are alternative methods for doing secondary-structure prediction or threading. The threading methods try to fit your sequence into a known 3-D structure. (We describe threading methods later in this chapter.)

6. **Click the Submit/Run Prediction button.**

TIP

Although it may seem convenient to submit all these analyses at once, our experience is that it is quicker and less confusing to go to the relevant sites and use them without the META intermediary.

Finding the best server around

TIP

There are many servers around the world and, as is to be expected, many use different kind of algorithms — which usually means that some perform faster than others. Of course, their authors keep updating them. If you want to know which one is currently the best server, you can check out EVA, the secondary-structure server monitoring system, at `cubic.bioc.columbia.edu/eva/`. (EVA is short for *EV*aluation of *A*utomatic protein structure prediction.)

From the Primary Structure to the 3-D Structure

In the previous section, we show you how to get some preliminary information about what your protein sequence may look like when it's folded into a biologically active molecule. Such information would involve helical and extended regions, loops, residues exposed at the surface, and so on.

If you have nothing else at hand, such information may be useful. For instance, it may help you predict the potential effect of a mutation — or choose the right protein fragment to make antibodies. Unfortunately, you'll soon find out that secondary structure predictions are far less useful than we would like them to be. They fall short of being the real thing: the detailed spatial representation of your molecule. In this section, we show you how to gather and use 3-D information to better understand what goes on in your protein sequence.

The good news for biologists is that plenty of experimental 3-D structure information is available on the Internet. As was the case when the molecular biologists of the world agreed to centralize their sequence data into the GenBank/EMBL/DDBJ databank repository, all structural biologists have agreed to deposit their 3-D structure coordinates into a single database: the Protein Data Bank. Everybody refers to this database by its acronym: PDB.

Retrieving and displaying a 3-D structure from a PDB site

Like other data repositories, the Protein Data Bank (PDB) offers a rather daunting interface that wasn't particularly designed with the nonspecialist in mind. Yet, in those rare cases where you know precisely what you're looking for — and even know what you're doing! — you may want to retrieve a protein 3-D structure dataset directly from one of the PDB sites. Before you query the PDB, be sure to collect some precise information about the structure you're looking for — such as the exact protein name or (even better) its PDB identifier. You can usually obtain this identifier from such user-friendly sources as the ExPASy/Swiss-Prot server or by using the various NCBI query tools. (See Chapter 4 for more on how to use these tools.)

Here's how to obtain and display a PDB structure. For now, let's assume that we are looking for the structure of an *Escherichia coli* (*E. coli*) protein named TolB, with PDB ID code 1CRZ.

1. **Point your browser to** `www.rcsb.org/pdb/`.

   This takes you to the PDB home page.

2. **Enter the PDB ID code** 1CRZ **in the search box in the middle at the very top of the page.**

3. **Click the adjacent SEARCH button.**

   Figure 11-3 shows the resulting output. This one-page Structure Summary form presents the essential information on this protein structure — including a small graphic, a bibliographical reference, its function, its organism of origin and — for the brave — some more obscure crystallographic parameters. You can get additional details on various aspects of this protein by clicking the various tags at the top of the form (see Figure 11-3).

For now, we only want to see what the protein molecule looks like in 3-D.

4. **Click the All Images link in the Display Options below the structure image (refer to Figure 11-3).**

   A new page appears, with an enlarged still view of the structure on the top, and an interactive view (a Java applet) at the bottom.

5. **Right-click the Still Image (in color).**

   A menu of options now allows you to either print the image, copy it, or save it to your own hard drive. It is now suitable for inclusion in reports or presentations. (Including such a figure is a great way to impress your boss or your professor!)

   But wait, there's more! You can actually get a better feeling for the shape of this protein using the bottom gray-scale Interactive View as follows:

6. **Right-click the Interactive View picture.**

   You are offered a large menu of viewing options. Explore them to see which functions of this Java applet actually work on your specific system. For instance, selecting Spin On puts the molecule in an endless rotation. (To stop the merry-go-round, select Spin Off.) This is a great way to gain a better understanding of the 3D shape of the molecule.

7. **Left-click and drag from any position in the structure, as shown in Figure 11-4.**

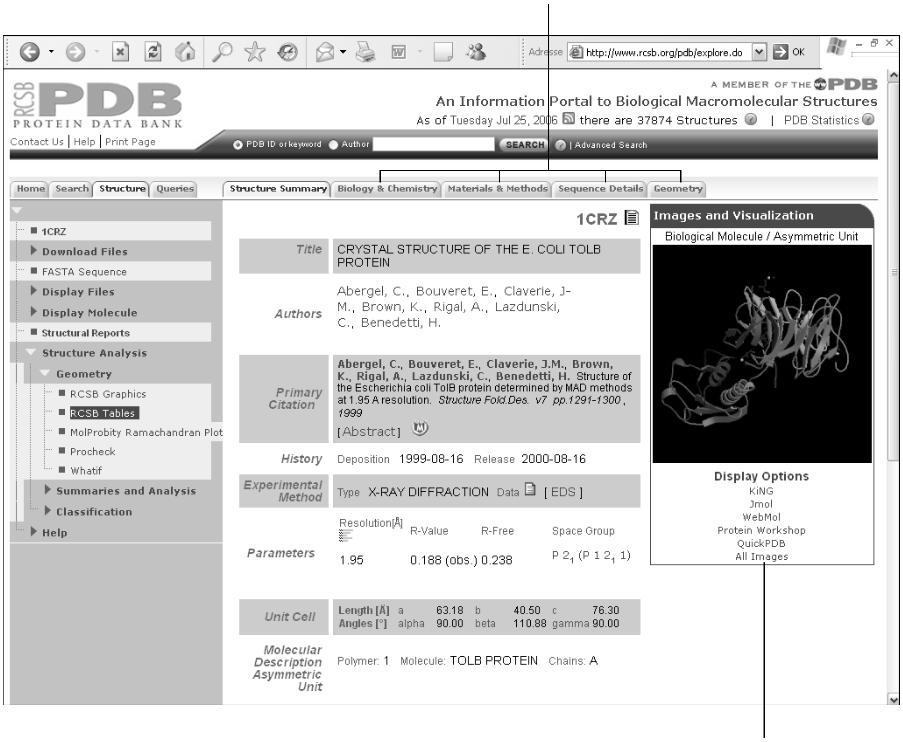
   This sets the molecule in controlled motion. Use this mode to inspect specific structural regions at your own pace, from any angle that you choose. (You can zoom in as well.) In addition, double-clicking allows you to select a given position from which you can measure distances and angles. Try it out for yourself to see what you can do; it's fun, and you may impress your colleagues with your brand new expertise in structural biology.

TIP

Feel free to play with the other display options (KiNG, WebMol, . . ., etc.) and see which one works best for you. See the section "Exploring the sequence/PDB structure relationship the interactive way," later in this chapter, for a tutorial on a more advanced tool that requires downloading and installing the CnD3 free software on your machine.

The preceding steps list shows you how to jump directly from the PDB entry to a visual representation of the molecule. This may not be enough, and you may also want to keep the PDB entry for further study on your own computer. Read on to find out how.

Click here for details.



Adresse http://www.rcsb.org/pdb/explore.do OK
RCSB PDB PROTEIN DATA BANK
A MEMBER OF THE PDB
Contact Us | Help | Print Page
○ PDB ID or keyword ● Author SEARCH | Advanced Search
Home Search Structure Queries
Structure Summary Biology & Chemistry Materials & Methods Sequence Details Geometry
1CRZ
Download Files
FASTA Sequence
Display Files
Display Molecule
Structural Reports
Structure Analysis
Geometry
RCSB Graphics
RCSB Tables
MolProbity Ramachandran Plot
Procheck
Whatif
Summaries and Analysis
Classification
Help
Images and Visualization
Biological Molecule / Asymmetric Unit
Display Options
KiNG
Jmol
WebMol
Protein Workshop
QuickPDB
All Images

An Information Portal to Biological Macromolecular Structures

As of Tuesday Jul 25, 2006 there are 37874 Structures | PDB Statistics

1CRZ

Title	CRYSTAL STRUCTURE OF THE E. COLI TOLB PROTEIN
Authors	Abergel, C., Bouveret, E., Claverie, J-M., Brown, K., Rigal, A., Lazdunski, C., Benedetti, H.
Primary Citation	**Abergel, C., Bouveret, E., Claverie, J.M., Brown, K., Rigal, A., Lazdunski, C., Benedetti, H.** Structure of the Escherichia coli TolB protein determined by MAD methods at 1.95 A resolution. *Structure Fold.Des. v7 pp.1291-1300, 1999* [Abstract]
History	Deposition 1999-08-16 Release 2000-08-16
Experimental Method	Type X-RAY DIFFRACTION Data [EDS]

Parameters	Resolution[Å]	R-Value	R-Free	Space Group
	1.95	0.188 (obs.)	0.238	P 2_1 (P 1 2_1 1)

Unit Cell						
Length [Å]	a	63.18	b	40.50	c	76.30
Angles [°]	alpha	90.00	beta	110.88	gamma	90.00

Molecular Description Asymmetric Unit: Polymer: 1 Molecule: TOLB PROTEIN Chains: A

Click here for an interactive view.

Figure 11-3: The PDB structure summary output form for query 1CRZ.

PDB files aren't meant to be read by nonspecialists. They are rather long because they contain a vast amount of data. For instance, the 1CRZ contains the detailed x, y, z coordinates of 3,488 atoms — as well as additional information about their connectivity!

If you're sure that you need to download a PDB file, you can simply stop after Step 3 in the previous step list — and look for the various links provided in the leftmost blue column of the screen. (Refer to Figure 11-3.) The possibilities are

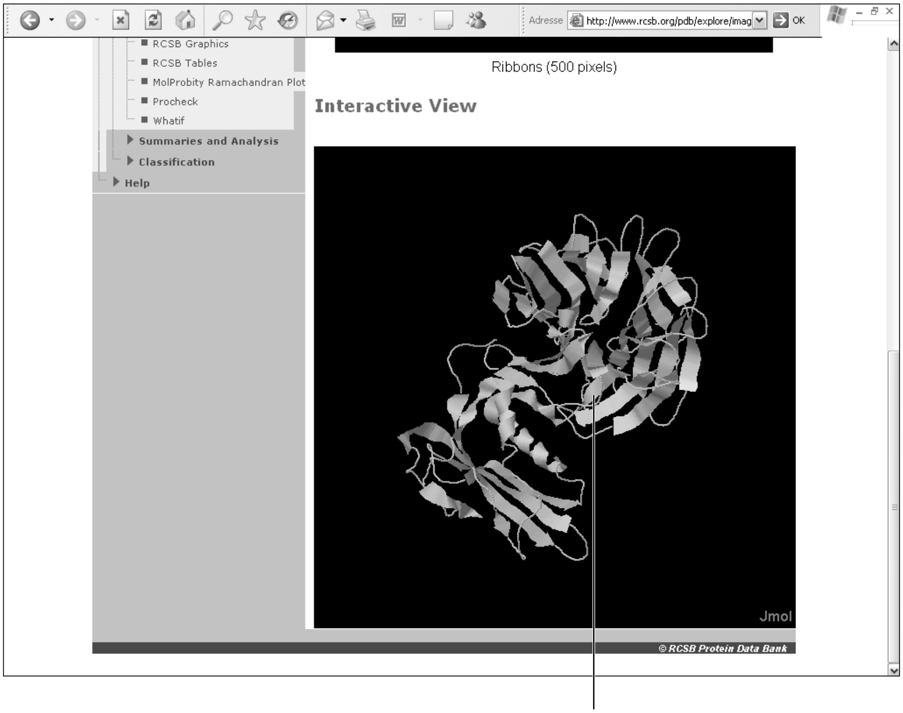
- Download Files (to save them to your own hard drive)
- Get the FASTA Sequence
- Display Files (to look at their internal format)
- Display Molecule
- Structural Reports

- Structure Analysis
- Help

Follow the first link (Download Files) to get your own copy of the coordinates. Feel free to explore the other links, in particular by clicking on the various tags at the top of the form. For instance, the Sequence Details link gives you a nice graphical picture of the actual secondary structure along with the sequence. You can compare it to the prediction that we made earlier in this chapter.

Guessing the 3-D structure of your protein

The previous section shows how to retrieve and quickly display a known 3-D–structure recorded in the PDB. This is definitely useful but falls short of addressing the questions we listed at the beginning of this chapter. Basically, we still do not know what the interesting bits of our own sequence look like in their 3-D context.



Click here to drag and rotate the molecule.

Figure 11-4: Interactive view of the 1CRZ PDB entry.

In the remaining sections of this chapter, we follow a realistic scenario to show you how you can sometimes get a fairly good idea of the 3-D structure of your protein *from its sequence alone* by doing a handful of simple operations.

We assume that you have determined the sequence and the secondary structure of your protein and that you are now asking this basic question: *How does it fold?* A simple way to answer this question is to look for a homologous protein with a known 3-D structure in the PDB. The next steps list shows you how.

Imagine, if you will, that we've just determined the sequence of the TolB gene of the bacteria *Rickettsia conorii* and we are curious about its structure. To satisfy our curiosity, we do the following:

1. **Fetch the protein sequence from NCBI:**

   a. Open a window in your browser and go to `www.ncbi.nlm.nih.gov`.

   b. Choose Protein from the Search drop-down menu.

   c. Type the identifier **NP_360043** in the query window and then click Go.

   d. When the answer comes back, change the Display format to FASTA.

   The *Rickettsia conorii* TolB protein sequence is now ready for use.

2. **Open a new browser window and point it to the NCBI BLAST server at** `www.ncbi.nlm.nih.gov/BLAST/`**.**

3. **Click the** Standard Protein-Protein BLAST [blastp] **link.**

   It is the first choice in the upper-right corner. The blastp input page appears.

4. **Select** pdb **from the Choose database drop-down menu.**

5. **Copy the *Rickettsia conorii* TolB sequence from the other browser window and paste it into the blastp query window.**

6. **Deselect the Do CD-search box (to simplify your output).**

7. **Click the BLAST! button.**

   An intermediate page appears, telling you that your request has been successfully submitted and put into the Blast Queue.

8. **Click the Format! button on the intermediate page and wait for the Blastp search to finish.**

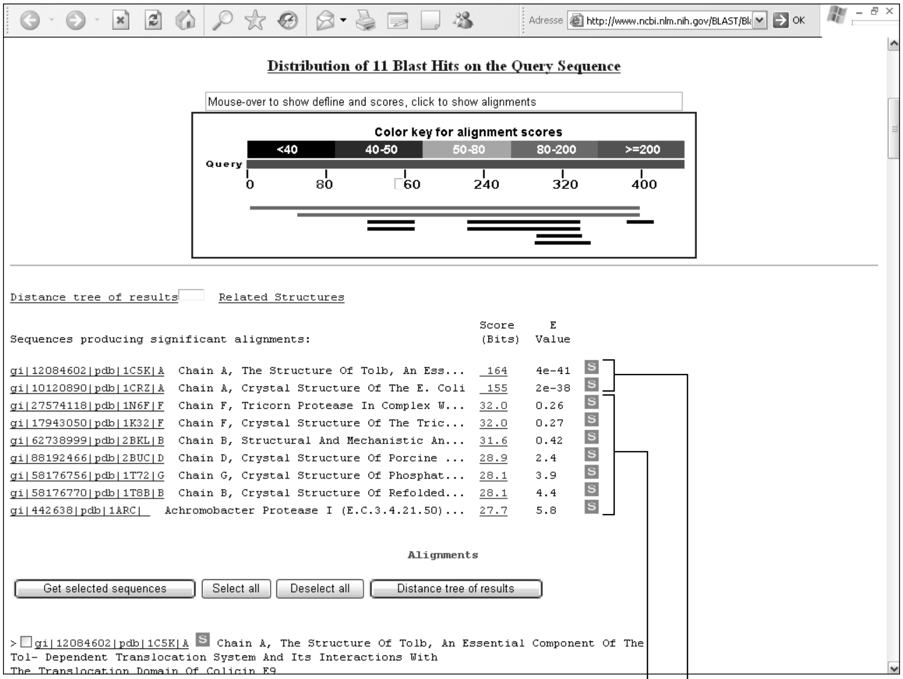
Figure 11-5 shows an example of the kinds of results you can expect. There are two strong matches with E-values much lower than 1. They are identified as

PDB entries 1C5K and 1CRZ. This is good news. If you click these matches, their identifiers reveal that they both correspond to the TolB protein from *E. coli* — the structure of which has been determined by two different laboratories.

This result is very interesting for two important reasons:

1. The homologous PDB sequences have 29 percent identical residues with our query.
2. The homology region covers the entire sequences.

By experience, we know that, given such a percentage of identical residues, the 3-D structure of our query protein (the *Rickettsia* TolB) is probably quasi-identical to the 3-D structure of its *E. coli* homologue. And so, we have a nearly perfect 3-D model to work with. The next section shows you how to use this model to further interpret the TolB protein sequence.



Adresse http://www.ncbi.nlm.nih.gov/BLAST/Bl
OK
Distribution of 11 Blast Hits on the Query Sequence
Mouse-over to show defline and scores, click to show alignments
Color key for alignment scores
<40
40-50
50-80
80-200
>=200
Query
0
80
160
240
320
400

```
Distance tree of results   Related Structures

                                                                   Score    E
Sequences producing significant alignments:                        (Bits)  Value

gi|12084602|pdb|1C5K|A  Chain A, The Structure Of Tolb, An Ess...   164    4e-41  S
gi|10120890|pdb|1CRZ|A  Chain A, Crystal Structure Of The E. Coli   155    2e-38  S
gi|27574118|pdb|1N6F|F  Chain F, Tricorn Protease In Complex W...   32.0   0.26   S
gi|17943050|pdb|1K32|F  Chain F, Crystal Structure Of The Tric...   32.0   0.27   S
gi|62738999|pdb|2BKL|B  Chain B, Structural And Mechanistic An...   31.6   0.42   S
gi|88192466|pdb|2BUC|D  Chain D, Crystal Structure Of Porcine ...   28.9   2.4    S
gi|58176756|pdb|1T72|G  Chain G, Crystal Structure Of Phosphat...   28.1   3.9    S
gi|58176770|pdb|1T8B|B  Chain B, Crystal Structure Of Refolded...   28.1   4.4    S
gi|442638|pdb|1ARC|   Achromobacter Protease I (E.C.3.4.21.50)...   27.7   5.8    S

                                   Alignments
```

```
> □gi|12084602|pdb|1C5K|A S Chain A, The Structure Of Tolb, An Essential Component Of The
Tol- Dependent Translocation System And Its Interactions With
The Translocation Domain Of Colicin E9
```

Figure 11-5: Blastp output of the Rickettsia conorii TolB sequence against PDB.

Non-significant matches

Two very good matches in PDB

Looking at sequence features in 3-D

On its own, a protein sequence is an important piece of information, but it becomes more revealing and valuable when you compare it with others from a variety of different species. Chapter 9 insists on the role of multiple alignments to identify the most significant regions of a sequence: Conserved residues (or, alternatively, highly variable ones) are often key to predicting or understanding a protein function. Going further, by precisely locating these conserved residues in space, it's possible to come up with additional clues as to their biological roles: Such residues, for example, can delineate a cavity at the active site of an enzyme — or, if they are found at the surface, one can assume they are good candidates for interactivity with other molecules, and so on.

We can use the example of the TolB protein family (of relatively unknown function) to illustrate the interplay between multiple alignments and structural analysis. Here's how it's done:

1. **First fetch several TolB homologue sequences from various bacterial species from NCBI:**

   a. Open a window in your browser and go to `www.ncbi.nlm.nih.gov`.

   b. Choose Protein from the drop-down Search menu.

   c. In the query window, type in the following identifiers: **NP_360043 (*R. conorii*)**, **NP_415268 (*E. coli*)**, **NP_404737 (*Y. pestis*)**, **NP_249663 (*P. aeruginosa*)**, and **NP_438543 (*H. influenzae*)**; then click Go.

   d. When the answer is returned, change the Display format to FASTA.

   e. Finally, use Send to Text to get rid of any parasitic characters.

   Five TolB protein sequences are now ready for use. Now we can build a multiple alignment out of these sequences.

2. **Open a new browser window and go to** `www.igs.cnrs-mrs.fr/Tcoffee/`**.**

3. **Click on Regular in the very top TCOFFEE option line.**

4. **Copy the five TolB sequences from the other browser window and paste them into the Tcoffee input window.**

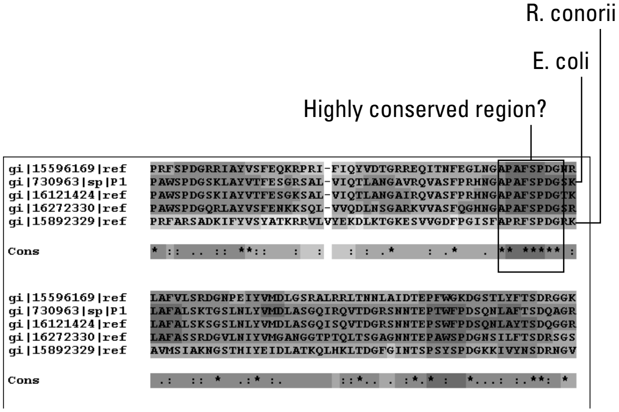
   Be sure to include their FASTA headers.

5. **Click Submit (at the bottom of the form).**

   The Tcoffee Results form automatically appears after an intermediary waiting page.

6. **Choose score_html from the Multiple Alignment Output menu.**

   This displays the multiple alignment in color with the most reliable regions in red. This part of the alignment is shown in Figure 11-6. It clearly reveals a segment of strictly conserved (invariant) residues shared by all these TolB proteins. Our challenge in the remaining pages of this chapter is to figure out where in the TolB structure these residues are located.



```
gi|15596169|ref     PRFSPDGRRIAYVSFEQKRPRI-FIQYVDTGRREQITNFEGLNGAPAFSPDGNR
gi|730963|sp|P1     PAWSPDGSKLAYVTFESGRSAL-VIQTLANGAVRQVASFPRHNGAPAFSPDGSK
gi|16121424|ref     PAWSPDGSKIAYVTFESGKSAL-VIQTLANGAIRQVASFPRHNGAPAFSPDGTK
gi|16272330|ref     PAWSPDGQRLAYVSFENKKSQL-VVQDLNSGARKVVASFQGHNGAPAFSPDGSR
gi|15892329|ref     PRFARSADKIFYVSYATKRRVLVYEKDLKTGKESVVGDFPGISFAPRFSPDGRK

Cons                * :: .. :: **::   :  :   : : .*    : .*   . ** ***** :


gi|15596169|ref     LAFVLSRDGNPEIYVMDLGSRALRRLTNNLAIDTEPFWGKDGSTLYFTSDRGGK
gi|730963|sp|P1     LAFALSKTGSLNLYVMDLASGQIRQVTDGRSNNTEPTWFPDSQNLAFTSDQAGR
gi|16121424|ref     LAFALSKSGSLNLYVMDLASGQISQVTDGRSNNTEPSWFPDSQNLAYTSDQGGR
gi|16272330|ref     LAFASSRDGVLNIYVMGANGGTPTQLTSGAGNNTEPAWSPDGNSILFTSDRSGS
gi|15892329|ref     AVMSIAKNGSTHIYEIDLATKQLHKLTDGFGINTSPSYSPDGKKIVYNSDRNGV

Cons                 .:  :: *  .:* :.       ::*.. . :*.* :  *...: :.**: *
```

Figure 11-6: An intriguing segment of strictly conserved residues in the TolB proteins.

Exploring the sequence/PDB structure relationship the interactive way

To interpret the significance of an invariant segment in our TolB proteins, we must be able to display its 3-D structure and *turn this structure around interactively* so we can look at it *from any angle.* Simultaneously, we must be able to refer to the protein sequence, and to *highlight* the residues we find interesting.

In short, we need a protein model that we can rotate around and analyze in parallel with its sequence. Not long ago, this kind of business was still the privilege of protein crystallographers working on expensive computers. Not anymore! If you can bear with us until the end of this chapter, you'll be doing it on your own PC in just a few minutes. Just follow our instructions:

1. **Point your browser to** `www.ncbi.nlm.nih.gov/Structure/`.

   The structure server of the NCBI appears.

2. **In the Search Entrez Structure/MMDB window at the top of the page, enter the PDB code of your relevant model. (MMDB stands for "*M*olecular *M*odeling *D*ata*B*ase.")**

   For our example, we'll use 1CRZ.

   TIP

   If you have the choice of several structures, a simple rule is to choose the one with the best resolution; for our example, the best resolution is 1CRZ (that is, 1.95 Angstroms). You'll find the Angstrom value in the ID card of each PDB entry (refer to Figure 11-3).

3. **Click the Go button.**

   Wait for a new page to appear.

4. **Click the 1CRZ link.**

   At this point, you should be looking at the MMDB structure card, as shown in Figure 11-7.

   This MMDB form exhibits a prominent View 3-D Structure button. What will happen next depends on the software already installed on your computer.

5. **Click the View 3-D Structure button.**

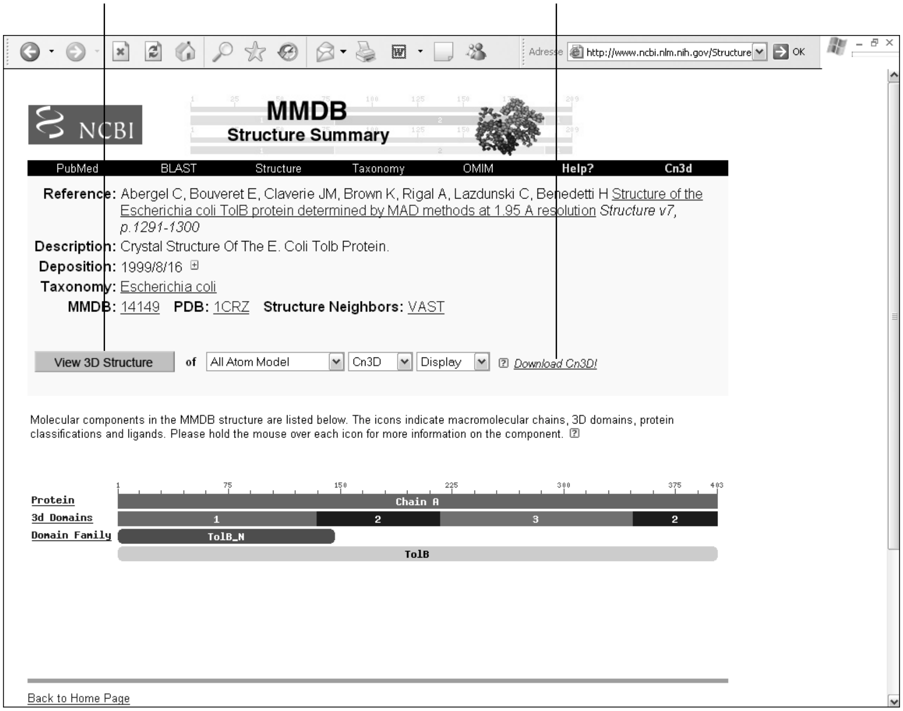
   Doing so downloads a file with the atomic coordinates of the 1CRZ protein structure. This file is formatted for the 3-D structure viewer Cn3D program. If nobody ever used your computer to display a protein structure in 3-D, a dialog box will pop up, asking you what to do with this file. It's telling you that your system doesn't contain the Cn3D program required to display the structure. In order to continue, you have to install the program. Here's how:

   a. Click the Download Cn3D! link.

      This takes you to the Cn3D home page, as shown in Figure 11-8. From here you can access a complete tutorial and/or continue the installation process.

Click here to view.

Click here to install Cn3D.



Reference: Abergel C, Bouveret E, Claverie JM, Brown K, Rigal A, Lazdunski C, Benedetti H Structure of the Escherichia coli TolB protein determined by MAD methods at 1.95 A resolution *Structure v7, p.1291-1300*

Description: Crystal Structure Of The E. Coli Tolb Protein.

Deposition: 1999/8/16

Taxonomy: Escherichia coli

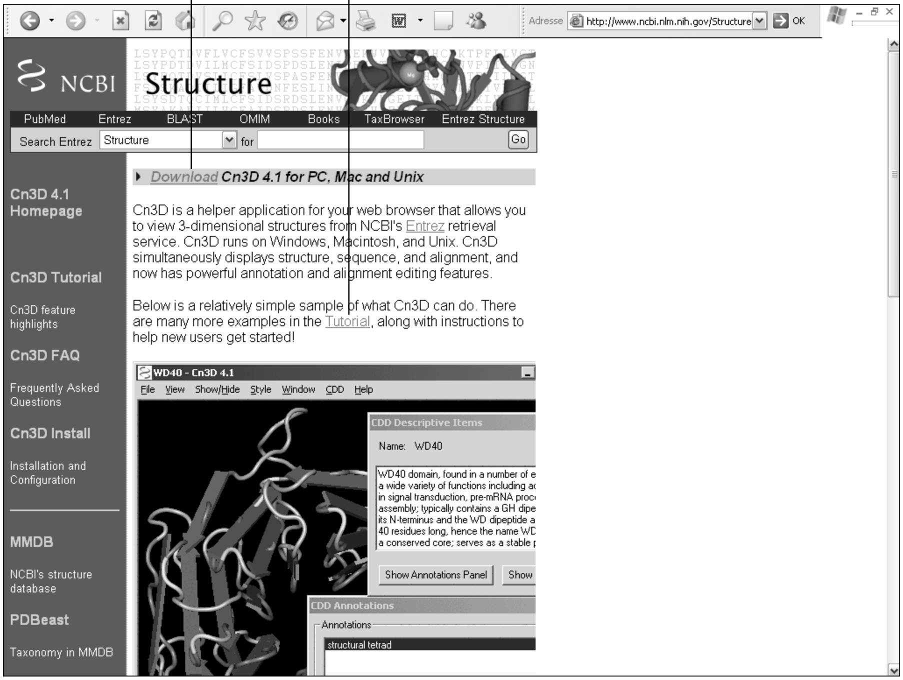
MMDB: 14149 **PDB:** 1CRZ **Structure Neighbors:** VAST

Molecular components in the MMDB structure are listed below. The icons indicate macromolecular chains, 3D domains, protein classifications and ligands. Please hold the mouse over each icon for more information on the component.

Figure 11-7: MMDB structure card for PDB entry 1CRZ.

Click here to install Cn3D.

Learn how to use Cn3D.



Download Cn3D 4.1 for PC, Mac and Unix

Cn3D is a helper application for your web browser that allows you to view 3-dimensional structures from NCBI's Entrez retrieval service. Cn3D runs on Windows, Macintosh, and Unix. Cn3D simultaneously displays structure, sequence, and alignment, and now has powerful annotation and alignment editing features.

Below is a relatively simple sample of what Cn3D can do. There are many more examples in the Tutorial, along with instructions to help new users get started!

WD40 domain, found in a number of e
a wide variety of functions including ac
in signal transduction, pre-mRNA proc
assembly; typically contains a GH dipe
its N-terminus and the WD dipeptide a
40 residues long, hence the name WD
a conserved core; serves as a stable p

Figure 11-8: The Cn3D home page.

b. Click Download Cn3D for PC, Mac, or Unix link.

c. In the window that appears (see Figure 11-9), select the version suitable for your computer: PC, Mac, or Unix.

Whatever platform you select, a new page appears with an FTP address, similar to the Windows page shown in Figure 11-10.

d. Click the FTP address link.

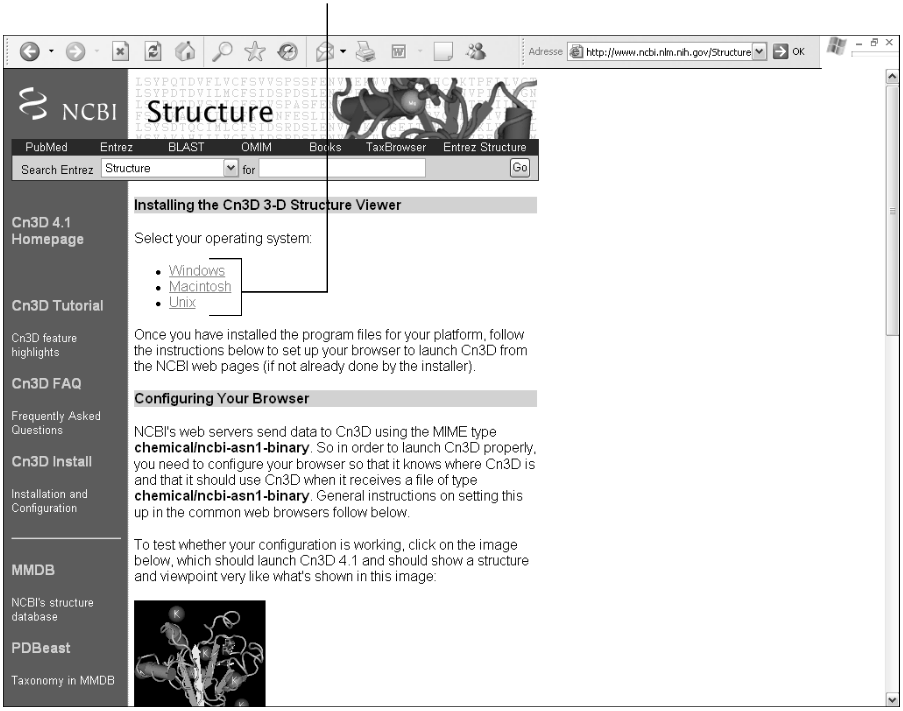
For our Windows system, the link is

```
ftp://ftp.ncbi.nih.gov/cn3d/Cn3D-4.1.msi
```

This downloads the CnD3 installer, which takes just a few seconds.

WARNING!

The system may also require you to install a newer version of the Windows Installer. Just follow the instructions. This also only takes a few extra seconds.



Select your system.

Installing the Cn3D 3-D Structure Viewer

Select your operating system:

- Windows
- Macintosh
- Unix

Once you have installed the program files for your platform, follow the instructions below to set up your browser to launch Cn3D from the NCBI web pages (if not already done by the installer).

Configuring Your Browser

NCBI's web servers send data to Cn3D using the MIME type **chemical/ncbi-asn1-binary**. So in order to launch Cn3D properly, you need to configure your browser so that it knows where Cn3D is and that it should use Cn3D when it receives a file of type **chemical/ncbi-asn1-binary**. General instructions on setting this up in the common web browsers follow below.

To test whether your configuration is working, click on the image below, which should launch Cn3D 4.1 and should show a structure and viewpoint very like what's shown in this image:

Figure 11-9: Installing Cn3D: System selection.

e. After everything is done, locate the Cn3D-4.1.msi icon on your desktop and click it twice.

   The Cn3D installation wizard welcomes you.

f. Follow its instructions and fill out the forms.

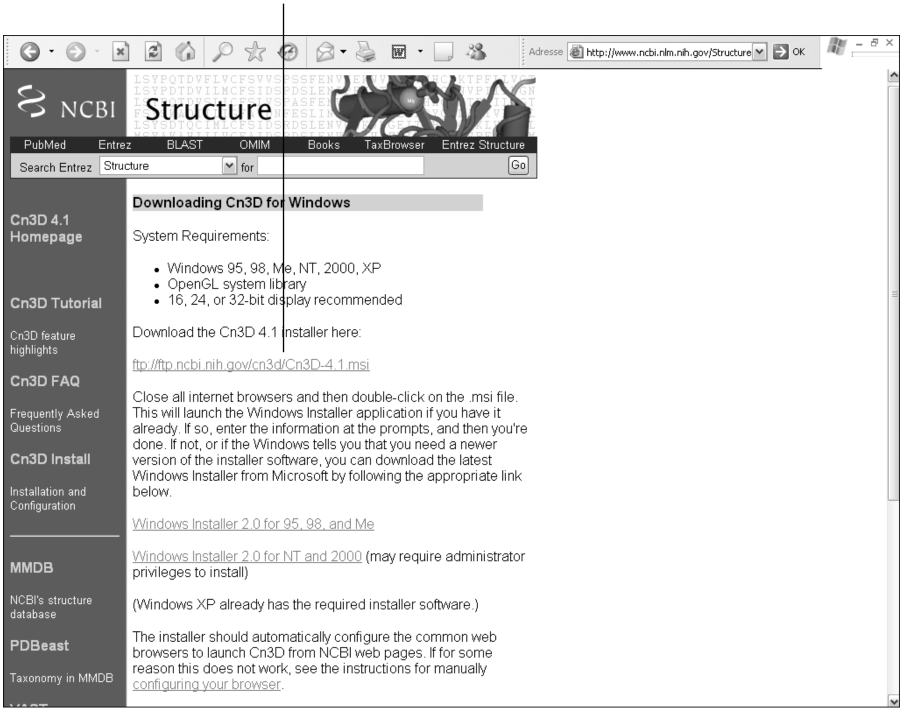
   You have to provide a name, an organization, and select a destination folder (default is fine). After the last step, the wizard confirms the successful installation.

Now that Cn3D is in your system, go back to NCBI and redo the sequence of steps that we describe in the previous section.

6. **Click the View 3D Structure button again.**

   Now that Cn3d is installed, the miracle occurs, and your screen should look like Figure 11-11. At the top of the MMDB page, the 3-D protein structure that corresponds to the 1CRZ PDB entry is displayed in a live window.

Download the Cn3D installer.



Adresse http://www.ncbi.nlm.nih.gov/Structure OK
NCBI Structure
PubMed Entrez BLAST OMIM Books TaxBrowser Entrez Structure
Search Entrez Structure for Go
Cn3D 4.1 Homepage
Cn3D Tutorial
Cn3D feature highlights
Cn3D FAQ
Frequently Asked Questions
Cn3D Install
Installation and Configuration
MMDB
NCBI's structure database
PDBeast
Taxonomy in MMDB

Downloading Cn3D for Windows

System Requirements:

- Windows 95, 98, Me, NT, 2000, XP
- OpenGL system library
- 16, 24, or 32-bit display recommended

Download the Cn3D 4.1 installer here:

ftp://ftp.ncbi.nih.gov/cn3d/Cn3D-4.1.msi

Close all internet browsers and then double-click on the .msi file. This will launch the Windows Installer application if you have it already. If so, enter the information at the prompts, and then you're done. If not, or if the Windows tells you that you need a newer version of the installer software, you can download the latest Windows Installer from Microsoft by following the appropriate link below.

Windows Installer 2.0 for 95, 98, and Me

Windows Installer 2.0 for NT and 2000 (may require administrator privileges to install)

(Windows XP already has the required installer software.)

The installer should automatically configure the common web browsers to launch Cn3D from NCBI web pages. If for some reason this does not work, see the instructions for manually configuring your browser.

Figure 11-10: Installing Cn3D: Installing the installer.

7. **Spin your protein around.**

   You can turn your protein around simply by moving your mouse; in fact, you can spin that protein around like a partner when you're dancing salsa! This is impressive, fun, and tremendously useful!

8. **Prepare your protein for taking a picture.**

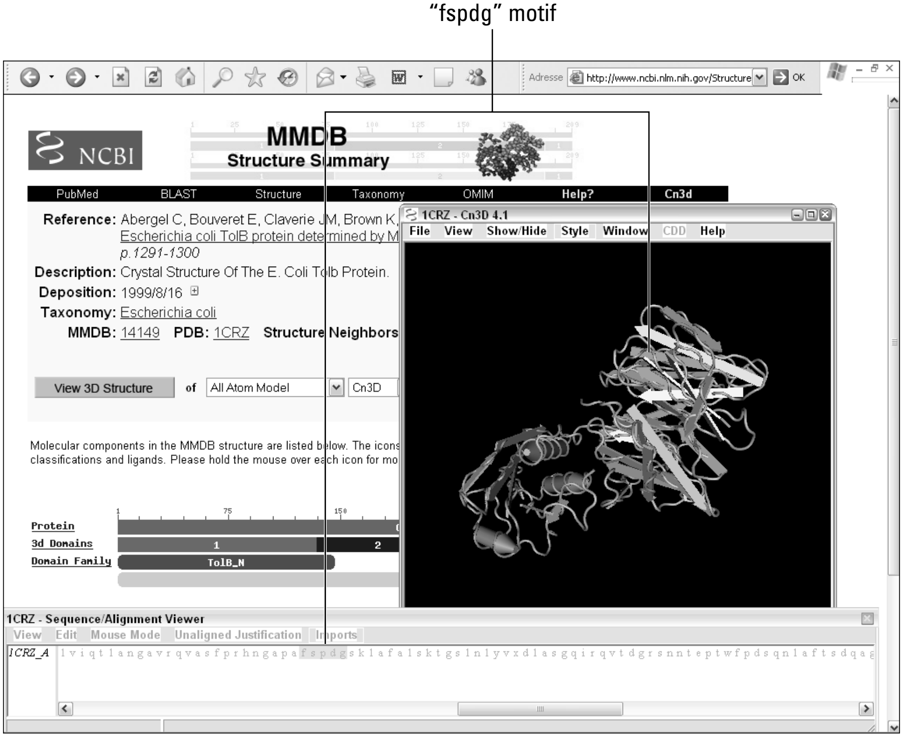
   Zoom in and out with the drop-down menus on the top bar of the window.

   Change the appearance of the model (wire, tube, ball and stick, or even solid space filling).

9. **Save a picture by choosing File➪Export PNG from the main menu in the Cn3D window.**

   This allows you to save a picture of the protein structure after choosing its appearance and orientation according to your personal preferences.

This spectacular show runs locally on your computer, independently of your Internet connection. Terminate your browser, and the live molecule stays on your screen! Neat! It also means that your data doesn't travel over the Internet. (This is important if you work for a paranoid boss.)



Reference: Abergel C, Bouveret E, Claverie JM, Brown K, Escherichia coli TolB protein determined by M p.1291-1300

Description: Crystal Structure Of The E. Coli Tolb Protein.

Deposition: 1999/8/16

Taxonomy: Escherichia coli

MMDB: 14149 PDB: 1CRZ Structure Neighbors

Molecular components in the MMDB structure are listed below. The icons classifications and ligands. Please hold the mouse over each icon for mo

```
1CRZ_A  1viqtlangavrqvasfprhngapafspdgsklafalsktgslnlyvxdlasgqirqvtdgrsnnteptwfpdsqnlaftsdqag
```

Figure 11-11: Interactive 3-D display of the E. coli TolB protein sequence and structure.

Analyzing the relationship between sequence and structure interactively

Last but not least, you probably noticed another window displayed below the live molecule. (Refer to Figure 11-11.) This is the sequence viewer. By default, this window displays the protein sequence associated with the PDB entry. Here we have the sequence of the *E. coli* TolB molecule.

The time has come to find out where in the structure that curious region of invariant residues that read "FSPDG" is located. With Cn3D, it could not be easier.

1. **In the sequence viewer top bar, click the View button.**
2. **From the pop-up menu that appears, choose the Find Pattern option.**
3. **In the pattern input box that pops up, type** FSPDG.

The corresponding segment is immediately highlighted in bright yellow in the sequence AND in the structure. To locate it in the sequence, scroll along the sequence by using the bottom cursor. To locate it in the structure, turn the molecule around.

You can click to select any residue or group of residues in the sequence viewer.

What can we tell about our invariant segment from all this? It's quite an interesting story, actually. Figure 11-11 shows that our segment occurs in a loop, well exposed at the surface of the molecule. Surface loops are notorious for being the most variable parts of proteins. (They are freer to mutate.) The strict conservation of these residues suggests that they may have a specific role in the protein function, such as participating in an interaction with another protein or a small molecule.

Beyond This Chapter

In this chapter, we've provided you with a very general idea of what biologists do when they analyze the structure of their proteins with their computers. The applications shown in this chapter probably cover 90 percent of what nonspecialists do when they study the relationships between their sequences and its potential structure.

However, we haven't even mentioned the many other tools and sophisticated approaches that exist for studying protein structures. So here we give you a quick list of what we have *not* talked about, with some related URLs if you're curious about delving deeper into 3-D structures.

Finding proteins with similar shapes

Following the determination of the 3-D structure of your preferred protein, you'll want to know if its shape is unique (what the specialists refer to as a "new fold") or if there are many like it. Just prepare a PDB file with the atomic coordinates of your protein and submit it to NCBI's structure-structure similarity search service (VAST) at `www.ncbi.nlm.nih.gov/Structure/VAST/vastsearch.html`.

Finding other PDB viewers

There are many ways to spin molecules around; you should be aware that, apart from Cn3D, other very powerful structure viewers are available over the Internet, such as

- **RasMol:** `www.rasmol.org`
- **DeepView, Swiss-PdbViewer:** `swissmodel.expasy.org/spdbv/`

Classifying your PDB structure

The classification of protein structures in categories of prototype folds and domain topologies is a hot topic. This is related to the problem of identifying and quantifying structural homologies between very distantly related proteins, as well as the threading problem (see the upcoming section "Threading sequences onto PDB structures"). The three main structure classification sites are

- **The CATH classification:** `www.cathdb.info`
- **The Dali 3D neighbor finding server:** `www.ebi.ac.uk/dali/`
- **The SCOP classification:** `scop.mrc-lmb.cam.ac.uk/scop/`

Doing homology modeling

The detailed modeling of the 3-D structure of a protein with significant sequence similarity to a homologous protein of known structure can go much beyond the simple example we have presented here. Specialists use energy minimization techniques to replace the residues in the original model and refine their positions. A couple of good entry points to find out more about this homology modeling approach are

- **Modeller:** `salilab.org/modeller/modeller.html`
- **SWISS-MODEL:** `swissmodel.expasy.org`

Folding proteins in a computer

Some groups are making constant progress in their attempts to generate a structural model for a protein solely from its sequence by simulating its folding process. This is referred to as the *ab initio* (or *in silico*) folding approach. The successes are still limited to a certain type of small protein sequences. Find out more about it at the Folding@home project Web site (`folding.stanford.edu/`).

Threading sequences onto PDB structures

Other groups have taken up the inverse problem of finding out which of the known structures might be best suited as a folding mold to a given protein sequence. This research field is known as *threading*. In the context of a growing database of structures, threading methods might perform increasingly well and expand the range of *homology modeling* to the case of sequences without any significant similarities in the PDB. To find out more about threading, check out these sites:

- **FUGUE:** `www-cryst.bioc.cam.ac.uk/~fugue/`
- **UCLA Fold recognition:** `www.doe-mbi.ucla.edu/Services/FOLD/`
- **The PROSPECT server:** `compbio.ornl.gov/structure/prospect/`

Looking at structures in movement

Some specialists are developing techniques to simulate the movements known to occur in protein structures, when stably folded. Understanding molecular dynamics is an important part of understanding the details of protein-protein interaction, or (for that matter) of enzymatic reactions. Find out more about molecular dynamics and protein movements at

- **The Brooks Lab site:** `brooks.Scripps.edu/`
- **The El Nemo site:** `www.igs.cnrs-mrs.fr/elnemo/`
- **The Database of Macromolecular Movements site:** `molmovdb.mbb.yale.edu/MolMovDB/`

Predicting interactions

Finally, researchers are tackling the difficult problem of predicting whether — and how — two different protein molecules (or a protein and a small molecule) might interact with each other. This is technically referred to as *molecular docking.* The top performance in this field is still very inadequate. In most cases, the 3-D structures of the complex of two interacting proteins cannot yet be computationally derived from their individual structures. Visit the following Web sites to find out more about docking:

- **AutoDock:** `www.scripps.edu/mb/olson/doc/autodock/`
- **FlexX:** `www.biosolveit.de/FlexX/`
- **FTDock:** `www.bmm.icnet.uk/docking/`
- **Hex:** `www.csd.abdn.ac.uk/hex/`
- **HotDock:** `wwwcs.uni-paderborn.de/~lst//HotDock/index.html`

Chapter 12

Working with RNA

In This Chapter

- Searching databases for RNA molecules
- Predicting the secondary structure of an RNA molecule
- Publishing an RNA secondary structure
- Finding non-coding RNAs over the Internet

> *Biology has at least 50 more interesting years.*
>
> — James D. Watson

In this chapter, we show you that it's possible to use a computer to analyze the most fascinating molecule of them all: the RNA molecule. RNA is incredible because it does almost anything you can think of — including transporting and transmitting genetic information (just like DNA) and performing catalytic functions (just like proteins). (The sidebar "RNA secondary structures" provides a closer look.) In fact, the more that biologists study RNA, the more they wonder why nature bothered inventing DNA and proteins!

There are sound arguments that support the belief that RNA is the true ancestral support of life, but we don't plan to go so far back in time in this chapter! We want to concentrate on another really exciting fact about RNA: It is one of the most clear-cut successes in bioinformatics. If you use them well, RNA computer-based methods can tell you as much as any million-dollar experiment, simply because RNA is much more predictable than DNA or proteins — and thus lends itself well to computer analysis.

This chapter is not here to give you an accurate view of the RNA world. This is simply too complex a subject. (You'll have to go back to your textbooks if you want the full story.) What we show here are the few things that you can easily do on a computer if you know what you're looking for and if you want to find out whether detailed RNA analysis could play a role in your project.

In the first section of this chapter, we show you how you can use mfold, a famous online server, to predict and display the secondary structure of your

RNA sequence. We also show you how to make sure that you actually *have* the correct structure.

The second section focuses on database search methods that are RNA-specific. If you know the secondary structure of an RNA molecule, you can use this information to scan databases for sequences that have a similar fold, even if their sequence similarity is very low. This is a very powerful technique that you can use to discover new members of an established gene class — or new members of an RNA family you've discovered yourself.

The last section gives you a very brief overview of one of the most active field of these last years: the study of small RNA molecules that are known as miRNAs or siRNAs. We show you resources you can use to look for these little guys and use them efficiently in your research.

Sadly, the big players, such as NCBI or the EBI, haven't put as much effort on the RNA side of things as they have in other fields. The good news is that the RNA community is very active and open. On the Internet, you can find a wealth of high-quality resources available to anyone who needs them, including specialized databases and powerful software. We introduce some of those resources in the last section of this chapter.

Predicting, Modeling, and Drawing RNA Secondary Structures

A major biological advance of the 1970s was the discovery that RNA can have complex 2-D and 3-D structures, as shown in Figure 12-1. The very exciting thing about these structures is that they obey laws that seem much simpler and much more predictable than the laws that reign in the world of proteins. Many of the RNA short- and long-range secondary structures rely on standard Watson and Crick base pairing, just as DNA does.

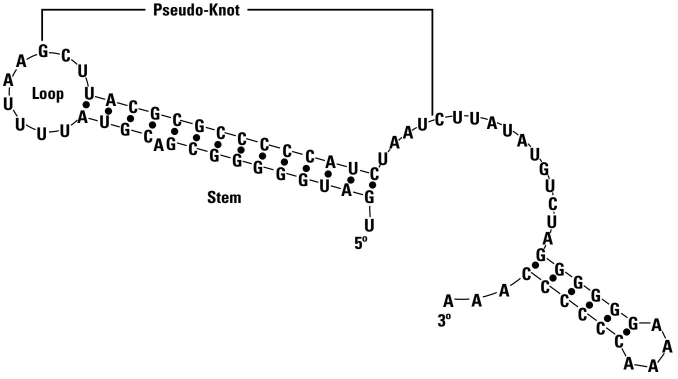


Figure 12-1: Typical secondary structures in RNA.

RNA secondary structures

A single-stranded RNA molecule works exactly like sticky tape. It will not be stable unless some of its exposed bases are protected from water. A good way to protect an RNA base from the solvent is to pair it with another RNA base. Of course, all possible base-pair combinations are not equally good. Pairing a guanine with a cytosine (for example) is more stabilizing than pairing an adenine with a uracile.

The pairing of these bases forms the *RNA secondary structure.* When a molecule contains two long stretches that are complementary, they yield a nice stable stem. (Refer to Figure 12-1.) The unpaired bases between the stem strands make up a loop. Stems don't have to be perfect; they can also contain unpaired residues (which RNA gurus name *bulges*).

We assume that the natural tendency of the RNA molecule is to reach its most stable conformation by assembling a nice collection of pairwise interactions, giving the molecule the highest stability it can have. This concept is what we call the *lowest-energy model.* Such a stable RNA structure always has a negative energetic value (such as –70 Kcal/mol); if you want to unfold it, you need to provide some energy (heat).

As is true for proteins, we don't know exactly how the RNA molecule finds this lowest energy form — but we know this happens very rapidly in the cell. The stability of an RNA fold doesn't completely depend on the number of GC pairs it contains. Many parameters — such as the stacking of base pairs (making some stems much more stable than others) and loop sizes — influence the fold. Sophisticated algorithms (like that found in the mfold program) can take these subtle effects into account when computing an optimal fold.

Tertiary interactions also play a role in overall stability. These include *pseudo-knots* (again, refer to Figure 12-1) that are usually long-range interactions between a loop and another portion of the RNA molecule. The interaction between the RNA molecule and other chemical elements — such as ions, proteins, or other RNAs — also plays an important role in its stabilization.

Unfortunately, we have difficulties predicting tertiary interactions or the effect of the proteins on the folding of an RNA molecule. So, when you predict the secondary structure of your RNA, this prediction usually depends on the assumption that this RNA folds on its own in the cell. Because this is almost never true, there is always a chance for your prediction to be partly or totally incorrect. The general rule is that the most energetically stable features tend to be reasonably close to the truth.

Using Mfold

Mfold is an implementation of the Zuker algorithm that makes it possible to predict the energetically optimal secondary structure of an RNA molecule. It uses a sophisticated model that can take into account many realistic physical parameters that affect the RNA folding — such as pH, temperature, and the local composition bias of your RNA.

Keep in mind that the energetic model that mfold uses ignores 3-D interactions or protein-RNA interactions that could stabilize a suboptimal fold. There is always a chance that a prediction from mfold does not correspond to the true biological fold of your RNA.

To help you make a decision, mfold returns suboptimal models so you can judge the stability of the optimal fold. It is also possible to force mfold to respect an interaction that you know is correct.

You cannot force pseudo-knots in mfold.

To try mfold, you'll need to do a couple of things. First and foremost, you'll need to fetch an RNA sequence. "Where?" you might ask. Well, some hunting is in order; when it comes to non-coding RNA, standard DNA databases are poorly equipped. A few specialized databases exist, but most of them are dedicated to ribosomal RNA. The last section of this chapter gives you the Internet address of a few of these resources (see Table 12-2).

What's the best length for your RNA sequence? Good question. The time mfold needs to analyze sequences increases very rapidly with their length. For instance, if it takes 1 second to fold a molecule that's 100 bases long, it takes about 10 seconds to fold a sequence that's 200 bases long.

The current limit of the server we're going to use for our example is 3,000 bases. This is convenient for most non-coding RNAs except the largest ribosomal ones.

With the preliminaries out of the way, it's time to do some mfolding. We're going to use a sample RNA sequence, but you're free to use one of your own if you want:

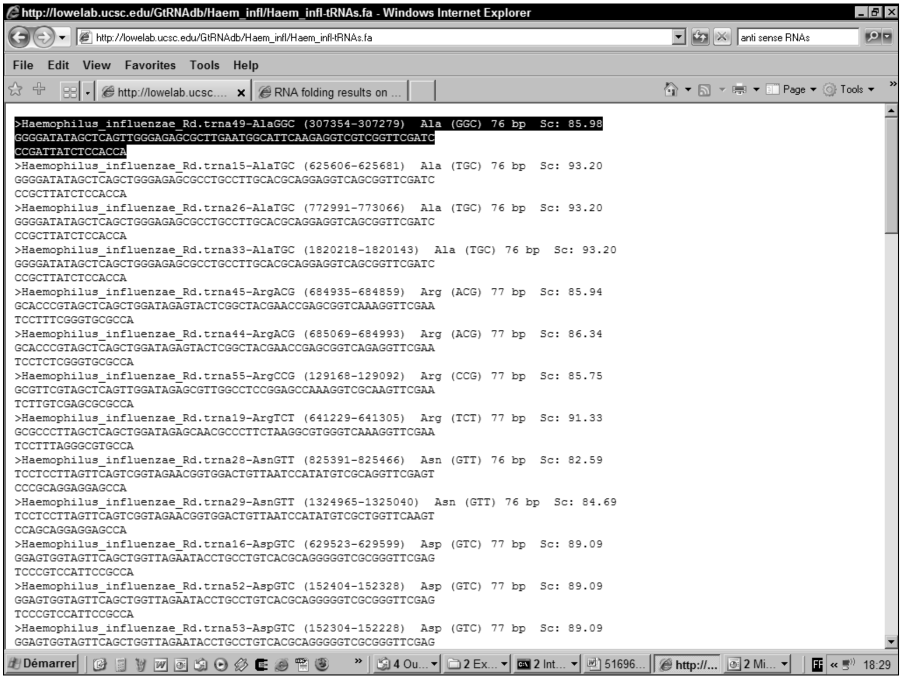
1. **Point your browser to** `lowelab.ucsc.edu/GtRNAdb/Haem_infl/Haem_infl-tRNAs.fa.`

   This returns a list of tRNA sequences predicted by Sean Eddy's tRNA scan server of the Haemophilus influenza genome. The sequences are ordered according to the reliability of their predictions. Sequences at the top of the list are almost 100 percent sure.

2. **Select the first tRNA sequence, as shown in Figure 12-2, and copy it to the Clipboard.**

3. **Point your browser to** `www.bioinfo.rpi.edu/applications/mfold/`**.**

   The Applications page of the Center for Bioinformatics (Rensselaer Polytechnic Institute) duly appears.


http://lowelab.ucsc.edu/GtRNAdb/Haem_infl/Haem_infl-tRNAs.fa - Windows Internet Explorer
http://lowelab.ucsc.edu/GtRNAdb/Haem_infl/Haem_infl-tRNAs.fa
anti sense RNAs
File Edit View Favorites Tools Help
http://lowelab.ucsc....
RNA folding results on ...
Page
Tools
Démarrer
4 Ou...
2 Ex...
2 Int...
51696...
http://...
2 Mi...
18:29

```
>Haemophilus_influenzae_Rd.trna49-AlaGGC (307354-307279)  Ala (GGC) 76 bp  Sc: 85.98
GGGGATATAGCTCAGTTGGGAGAGCGCTTGAATGGCATTCAAGAGGTCGTCGGTTCGATC
CCGATTATCTCCACCA
>Haemophilus_influenzae_Rd.trna15-AlaTGC (625606-625681)  Ala (TGC) 76 bp  Sc: 93.20
GGGGATATAGCTCAGCTGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAGCGGTTCGATC
CCGCTTATCTCCACCA
>Haemophilus_influenzae_Rd.trna26-AlaTGC (772991-773066)  Ala (TGC) 76 bp  Sc: 93.20
GGGGATATAGCTCAGCTGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAGCGGTTCGATC
CCGCTTATCTCCACCA
>Haemophilus_influenzae_Rd.trna33-AlaTGC (1820218-1820143)  Ala (TGC) 76 bp  Sc: 93.20
GGGGATATAGCTCAGCTGGGAGAGCGCCTGCCTTGCACGCAGGAGGTCAGCGGTTCGATC
CCGCTTATCTCCACCA
>Haemophilus_influenzae_Rd.trna45-ArgACG (684935-684859)  Arg (ACG) 77 bp  Sc: 85.94
GCACCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACCGAGCGGTCAAAGGTTCGAA
TCCTTTCGGGTGCGCCA
>Haemophilus_influenzae_Rd.trna44-ArgACG (685069-684993)  Arg (ACG) 77 bp  Sc: 86.34
GCACCCGTAGCTCAGCTGGATAGAGTACTCGGCTACGAACCGAGCGGTCAGAGGTTCGAA
TCCTCTCGGGTGCGCCA
>Haemophilus_influenzae_Rd.trna55-ArgCCG (129168-129092)  Arg (CCG) 77 bp  Sc: 85.75
GCGTTCGTAGCTCAGTTGGATAGAGCGTTGGCCTCCGGAGCCAAAGGTCGCAAGTTCGAA
TCTTGTCGAGCGCGCCA
>Haemophilus_influenzae_Rd.trna19-ArgTCT (641229-641305)  Arg (TCT) 77 bp  Sc: 91.33
GCGCCCTTAGCTCAGCTGGATAGAGCAACGCCCTTCTAAGGCGTGGGTCAAAGGTTCGAA
TCCTTTAGGGCGTGCCA
>Haemophilus_influenzae_Rd.trna28-AsnGTT (825391-825466)  Asn (GTT) 76 bp  Sc: 82.59
TCCTCCTTAGTTCAGTCGGTAGAACGGTGGACTGTTAATCCATATGTCGCAGGTTCGAGT
CCCGCAGGAGGAGCCA
>Haemophilus_influenzae_Rd.trna29-AsnGTT (1324965-1325040)  Asn (GTT) 76 bp  Sc: 84.69
TCCTCCTTAGTTCAGTCGGTAGAACGGTGGACTGTTAATCCATATGTCGCTGGTTCAAGT
CCAGCAGGAGGAGCCA
>Haemophilus_influenzae_Rd.trna16-AspGTC (629523-629599)  Asp (GTC) 77 bp  Sc: 89.09
GGAGTGGTAGTTCAGCTGGTTAGAATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAG
TCCCGTCCATTCCGCCA
>Haemophilus_influenzae_Rd.trna52-AspGTC (152404-152328)  Asp (GTC) 77 bp  Sc: 89.09
GGAGTGGTAGTTCAGCTGGTTAGAATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAG
TCCCGTCCATTCCGCCA
>Haemophilus_influenzae_Rd.trna53-AspGTC (152304-152228)  Asp (GTC) 77 bp  Sc: 89.09
GGAGTGGTAGTTCAGCTGGTTAGAATACCTGCCTGTCACGCAGGGGGTCGCGGGTTCGAG
```

Figure 12-2: Selection of a tRNA sequence.

As an alternative, you can also use one of the two European mirrors of mfold:

`bibiserv.techfak.uni-bielefeld.de/cgi-bin/mfold_submit`

`bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html`

or its Australian counterpart:

`mfold.burnet.edu.au`

4. **Scroll down the page and click the RNA Folding link.**

   The mfold server page appears, as (partially) shown in Figure 12-3.

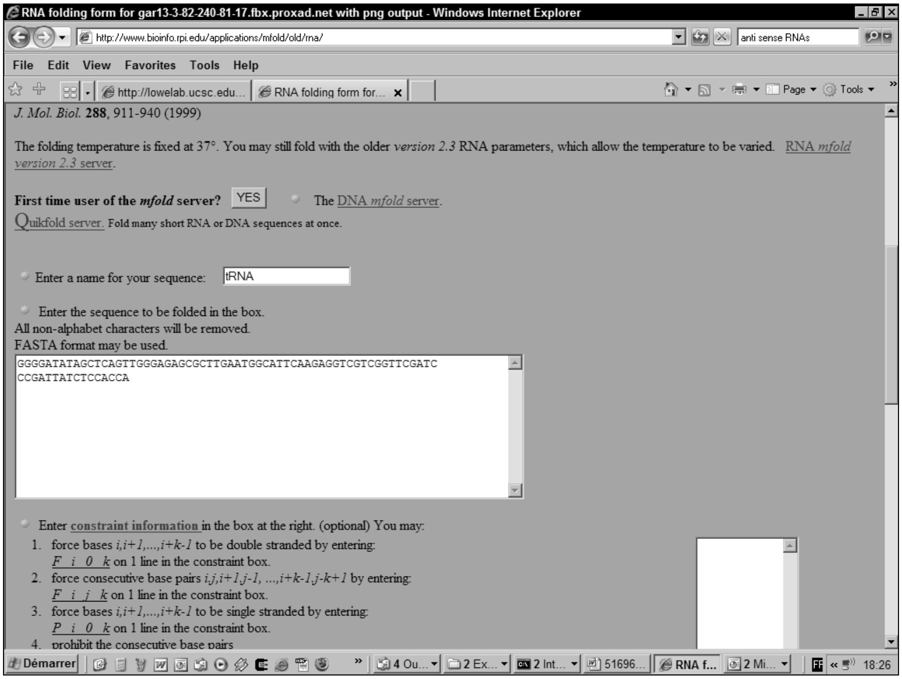
5. **Enter a title for your sequence in the Name field.**

   For our example, we named our sequence tRNA.

6. **Paste your tRNA sequence into the Sequence window.**

   You can safely keep all the other default parameters.

   See the next step list (in the "Forcing interaction in mfold" section) if you want to provide the program with some predefined constraints.


RNA folding form for gar13-3-82-240-81-17.fbx.proxad.net with png output - Windows Internet Explorer
http://www.bioinfo.rpi.edu/applications/mfold/old/rna/
anti sense RNAs
File Edit View Favorites Tools Help
http://lowelab.ucsc.edu...
RNA folding form for...
Page Tools
Démarrer
4 Ou...
2 Ex...
2 Int...
51696...
RNA f...
2 Mi...
18:26

J. Mol. Biol. **288**, 911-940 (1999)

The folding temperature is fixed at 37°. You may still fold with the older *version 2.3* RNA parameters, which allow the temperature to be varied. RNA *mfold* *version 2.3* server.

First time user of the *mfold* server? YES ○ The DNA *mfold* server.

Quikfold server. Fold many short RNA or DNA sequences at once.

○ Enter a name for your sequence: tRNA

○ Enter the sequence to be folded in the box.
All non-alphabet characters will be removed.
FASTA format may be used.

```
GGGGATATAGCTCAGTTGGGAGAGCGCTTGAATGGCATTCAAGAGGTCGTCGGTTCGATC
CCGATTATCTCCACCA
```

○ Enter constraint information in the box at the right. (optional) You may:

1. force bases *i,i+1,...,i+k-1* to be double stranded by entering:
   F i 0 k on 1 line in the constraint box.
2. force consecutive base pairs *i,j,i+1,j-1, ...,i+k-1,j-k+1* by entering:
   F i j k on 1 line in the constraint box.
3. force bases *i,i+1,...,i+k-1* to be single stranded by entering:
   P i 0 k on 1 line in the constraint box.
4. prohibit the consecutive base pairs

Figure 12-3: Mfold server home page.

7. **Scroll down the page and click the Fold RNA button near the bottom of the form.**

   The Output page (eventually) appears, as shown in Figure 12-4. The output itself looks rather complicated because it contains links to your results in almost any format you can think of. It also contains thermodynamic information concerning your model, so that you can tell how stable they are.

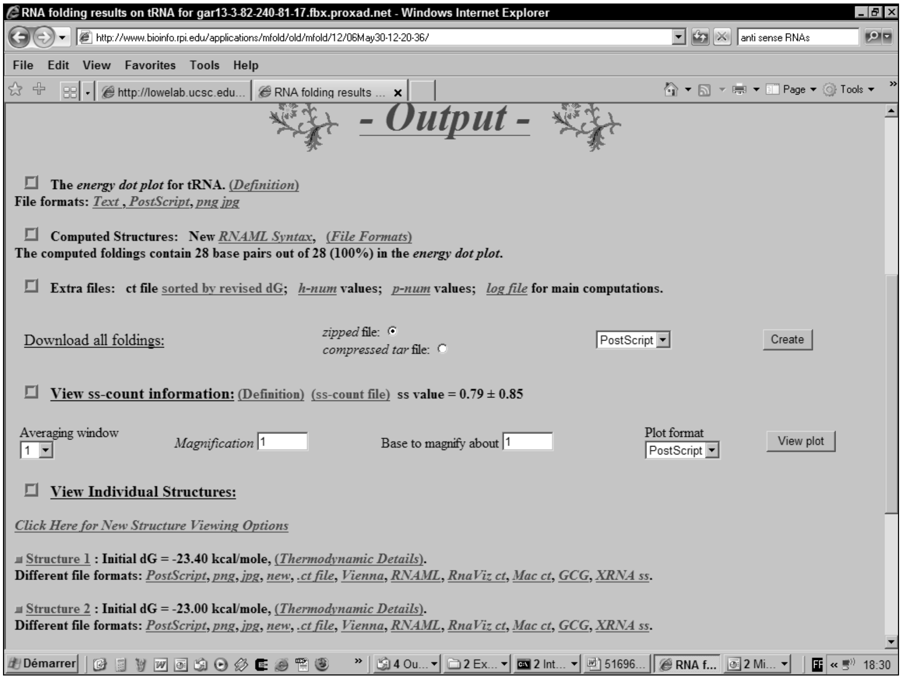
8. **Display the jpg picture of the best-scoring model.**

   a. Scroll down the Output page until you find the section titled View Individual Structures (refer to Figure 12-4).

      The structures are ranked according to their stability. For instance, Structure1 is the most stable model, Structure2 is the second best, and so on.

   b. Look for the listing of different file formats, and then click the jpeg link associated with the Structure1 link. This will load a new page, containing the most stable secondary structure of your sequence (Figure 12-5). This is not necessarily the correct one, but a good candidate.

9. **Save your secondary structure prediction by right-clicking on the image that has just appeared and then choosing the Save As option.**


RNA folding results on tRNA for gar13-3-82-240-81-17.fbx.proxad.net - Windows Internet Explorer
http://www.bioinfo.rpi.edu/applications/mfold/old/mfold/12/06May30-12-20-36/
anti sense RNAs
File Edit View Favorites Tools Help
http://lowelab.ucsc.edu...
RNA folding results ...
Page
Tools
Create
View plot
Démarrer
4 Ou...
2 Ex...
2 Int...
51696...
RNA f...
2 Mi...
18:30

- Output -

□ The *energy dot plot* for tRNA. (*Definition*)
File formats: *Text* , *PostScript*, *png jpg*

□ Computed Structures: New *RNAML Syntax*, (*File Formats*)
The computed foldings contain 28 base pairs out of 28 (100%) in the *energy dot plot*.

□ Extra files: ct file sorted by revised dG; *h-num* values; *p-num* values; *log file* for main computations.

Download all foldings: *zipped* file: ● *compressed tar* file: ○ PostScript

□ View ss-count information: (Definition) (ss-count file) ss value = 0.79 ± 0.85

Averaging window 1 *Magnification* 1 Base to magnify about 1 Plot format PostScript

□ View Individual Structures:

Click Here for New Structure Viewing Options

Structure 1 : Initial dG = -23.40 kcal/mole, (*Thermodynamic Details*).
Different file formats: *PostScript*, *png*, *jpg*, *new*, *.ct file*, *Vienna*, *RNAML*, *RnaViz ct*, *Mac ct*, *GCG*, *XRNA ss*.

Structure 2 : Initial dG = -23.00 kcal/mole, (*Thermodynamic Details*).
Different file formats: *PostScript*, *png*, *jpg*, *new*, *.ct file*, *Vienna*, *RNAML*, *RnaViz ct*, *Mac ct*, *GCG*, *XRNA ss*.

Figure 12-4: Output of the mfold server.

Interpreting mfold results

After you've given mfold your sequence, it looks for the arrangement that yields the secondary structures with the lowest possible energy. Of course, the correctness of this secondary structure depends on the correctness of the energy model; if the model is wrong, the structure is wrong. (See the "RNA secondary structures" sidebar.)

The general rule is that the more stable the conformation mfold predicts, the more likely this interaction is to be correct. To help you get an idea of the nature of this stability, mfold's Output page displays several results related to the stability of your RNA. Among these, you may want to give further attention to the following:

- **The Energy Dot Plot:** The energy dot plot shows the stems that are part of the optimal fold of your sequence.
  - A stem shows up as a black diagonal, perpendicular to the main diagonal.
  - Distal interaction (that is, pairs that involve nucleotides very far from each other) appears as diagonals close to the top-right (or bottom-left) corner.

- Colored diagonals indicate suboptimal secondary structures. The color ranges from red (the best) to yellow (the worst).

- **The View Individual Structures section:** This section contains data relative to the various structures mfold has computed. The optimal structure is the top one.

  This section also indicates the free energy of your structure. This energy must be as low as possible. For a molecule that's around 200 bases long, you should expect a free energy lower than –50 Kcal to consider the reported fold as plausible.

- **The Dot Plot Folding Comparisons section:** This very interesting section makes it possible to compare the fold of several suboptimal solutions. In order to do so, follow these steps:

  1. **Choose .jpg from the Image Format drop-down menu.**
  2. **From the Compare Selected Foldings line, select all the alternative structures you want to compare.**
  3. **Click the Do the Comparison button, as shown in Figure 12-6.**

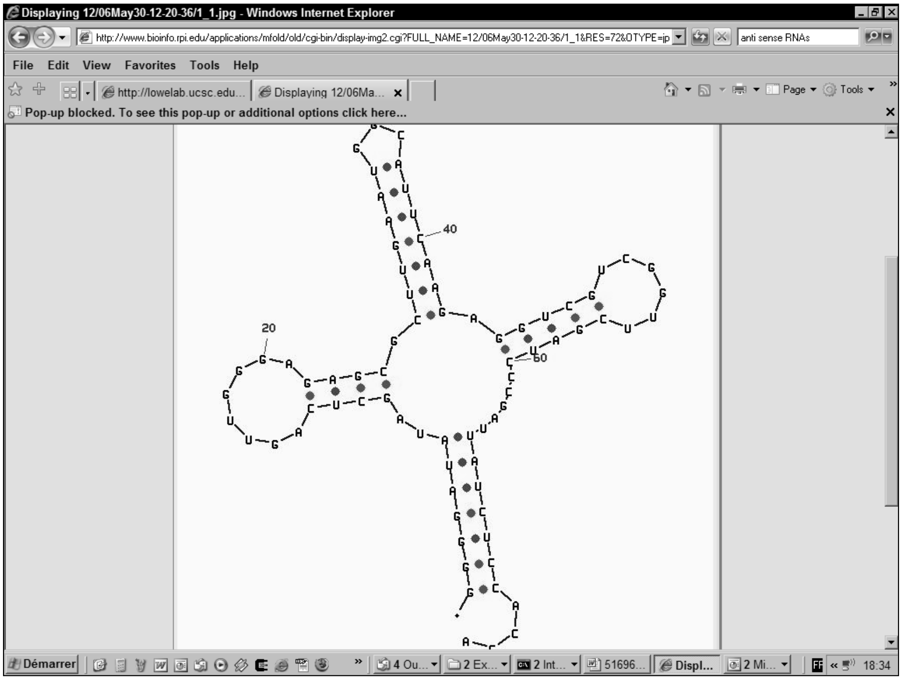
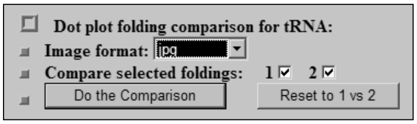

Displaying 12/06May30-12-20-36/1_1.jpg - Windows Internet Explorer
http://www.bioinfo.rpi.edu/applications/mfold/old/cgi-bin/display-img2.cgi?FULL_NAME=12/06May30-12-20-36/1_1&RES=72&OTYPE=jp
anti sense RNAs
File Edit View Favorites Tools Help
http://lowelab.ucsc.edu...
Displaying 12/06Ma...
Page
Tools
Pop-up blocked. To see this pop-up or additional options click here...
40
20
60
Démarrer
4 Ou...
2 Ex...
2 Int...
51696...
Displ...
2 Mi...
18:34

Figure 12-5: Representation of an RNA secondary structure prediction.



□ Dot plot folding comparison for tRNA:

Image format: jpg

Compare selected foldings: 1 ☑ 2 ☑

Figure 12-6: Comparing alternative secondary structures with mfold.

A graph similar to the one displayed in Figure 12-5 appears. It indicates the elements of predicted secondary structures that are common to the selected suboptimal solutions.

Colored elements correspond to stems that occur only in suboptimal secondary structure predictions.

Finding a high consistency between the optimal solution and the suboptimal ones makes it more likely that the optimal solution is, in fact, biologically correct.

The predictions you make on a single sequence can yield interesting results. RNA is at its best, however, when you make the prediction on a multiple alignment rather than on a single sequence, thanks to what is known as the *covariance phenomenon*.

Unfortunately, multiple-sequence-based predictions are not available for free over the Internet, but if you are ready to install some software — and you have a powerful machine — you can use a handy bit of software called Cove, available for download at `www.genetics.wustl.edu/eddy/software/#cove`.

or some of the methods available through registration in Robin Gutell's laboratory at `www.rna.icmb.utexas.edu`.

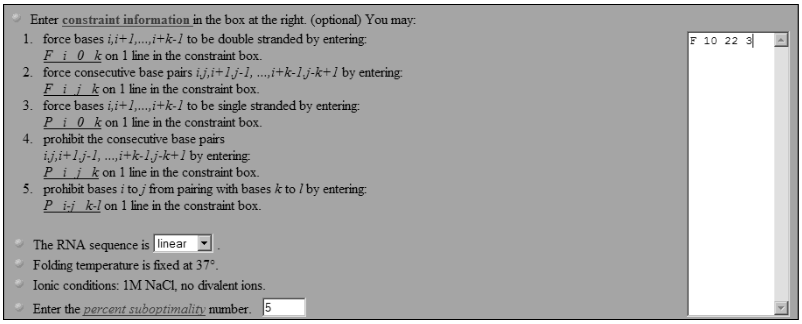
Forcing interaction in mfold

If you have some experimental data and you know that specific base pairs occur in your RNA sequence, you can force mfold to use this information. For example, you can tell mfold that nucleotides 10, 11, and 12 interact respectively with nucleotides 22, 21, and 20. To do this, you write a line of instruction in the interaction form of mfold that says

```
F, 10, 22, 3
```

This means that you (F)orce an interaction to start with nucleotide 10 — paired with nucleotide 22 — and that you want this interaction to be the beginning of a stem that is 3 nucleotide pairs long (nucleotide 10 with 22, 11 with 21, and 12 with 20).

Figure 12-7 shows you what the interaction form of mfold looks like when it is filled out correctly.



- Enter constraint information in the box at the right. (optional) You may:
  1. force bases *i,i+1,...,i+k-1* to be double stranded by entering:
     F i 0 k on 1 line in the constraint box.
  2. force consecutive base pairs *i,j,i+1,j-1, ...,i+k-1,j-k+1* by entering:
     F i j k on 1 line in the constraint box.
  3. force bases *i,i+1,...,i+k-1* to be single stranded by entering:
     P i 0 k on 1 line in the constraint box.
  4. prohibit the consecutive base pairs
     i,j,i+1,j-1, ...,i+k-1,j-k+1 by entering:
     P i j k on 1 line in the constraint box.
  5. prohibit bases *i* to *j* from pairing with bases *k* to *l* by entering:
     P i-j k-l on 1 line in the constraint box.

```
F 10 22 3
```

- The RNA sequence is linear ▾ .
- Folding temperature is fixed at 37°.
- Ionic conditions: 1M NaCl, no divalent ions.
- Enter the *percent suboptimality* number. 5

Figure 12-7: Forcing secondary interaction in mfold.

Searching Databases and Genomes for RNA Sequences

In a cell, all genes are transcribed into RNA molecules so the cell can use them. One thing we all forget sometimes is that there are two kinds of RNAs: those that the cell turns into a protein (using the ribosome machinery) and those the cell uses directly. The latter are known as *non-coding RNAs.*

Non-coding RNAs constitute a rich family, with new members discovered every year. The more we understand the basic mechanisms that govern the life of a cell, the more we realize that non-coding RNAs are everywhere. They help with transcribing, splicing, replicating, and probably many other things we aren't yet aware of. It seems that every basic mechanism of life heavily relies on these little guys — and arguing that a big chunk of biology's future belongs to them is probably the safest bet we'll make in this book.

Unfortunately, with the exception of the ribosomal RNA, these small non-coding RNAs tend to be rather short and poorly conserved. Their most common feature is often their secondary structure. This makes it difficult to find them using any of the homology-based methods we routinely use for finding proteins. (For more on such methods, see Chapter 7.)

The methods we show you in this section initiate database searches using not only the sequence of RNAs but also information related to the known secondary structures. If you have just used mfold to predict the fold of the RNA family you're interested in, you can now use the fold you've discovered to look for new members of the family. In this section, we first show you how to use tRNAscan to look only for tRNAs, and we also introduce PatScan, a method that allows you to look for RNAs with a secondary structure that you can specify yourself.

Finding tRNAs in a genome

tRNAs are small non-coding RNAs that the cell uses to assemble proteins. They constitute one of the best examples of important non-coding genes that are difficult to localize in a genome. All the bacterial and eukaryotic genomes contain tRNAs. Unfortunately, these ubiquitous genes are just as difficult to identify as any small non-coding RNA. They require database search techniques much more sophisticated than BLAST.

The good news here is that there are some very efficient programs out there for predicting tRNA genes in a eukaryotic or a prokaryotic genome. The state-of-the-art method is RNAscan-SE that you can access from the server at Washington University in St. Louis: `selab.janelia.org/tRNAscan-SE/`.

Using PatScan to look for RNA patterns

tRNAscan is perfect if you're interested only in tRNAs. However, if your interest lies in a specific RNA family, you want to use a tool that's more general and enables you to use your own knowledge of the secondary structure of your RNA family. PatScan is ideal for this purpose.

PatScan lets you search databases with patterns that can accurately describe RNA secondary structures. From a computational point of view, this type of search is quite expensive, which is why PatScan requires your e-mail address and returns a result after a few hours. (In our experience, however, PatScan has never taken more than half an hour to return the result.)

If PatScan is not the route you want to go — perhaps because you don't want your sequences traveling across the Internet — you can use RNAmot, an alternative program that's similar to PatScan. To use the program, however, you'll need to install the software on your own computer. If you want to do this, be aware that this is not really a tool for novices. RNAmot is available for download at `www.esil.univ-mrs.fr/~dgaut/download/`.

Designing a PatScan pattern is the trickiest step when using PatScan. It is something that may require a bit of thought. If you want to know the exact rules that govern the PatScan patterns, you can read the excellent online documentation of this program, available at `www-unix.mcs.anl.gov/compbio/PatScan/HTML/matching_nucleotides.html`.

To help design your pattern, start by identifying the elements you're interested in. Use the following checklist as a guide:

- On the primary structure, report the secondary structure elements.
- Identify the corresponding stems and name them. The name of the left side of the first stem is *p1* and the name of its right side is *~p1*.
- Count the length of each of these elements.
- In each stem element, identify non-canonical Watson and Crick base pairs.
- Write this structure down with your own vocabulary.

  For instance:

  [stem p1 of length 8 to 9] [Loop of length 3 to 8] [stem ~p1 of length 8 to 9, complementary to the left stem p1]
- Turn your secondary structure into a PatScan Pattern.
- p1=8...9 defines the stem p1 that contains 8 to 9 nucleotides. [Note the 3 periods (...) between the 8 and 9.]
- 3...8 defines a pattern of 3 to 8 nucleotides.
- ~p1 defines a stem reverse complementary to p1.

  You can then write your pattern like this: p1=8...9 3...8 ~p1.

  If you want to define several stems, give them different names, such as p1, p2, p3, and so on.

  If you need to, you can have nested stems, like p1=8...9 3...8 p2=5...7 4...5 ~p2 3...8 ~p1.

- You can accommodate in PatScan almost anything you need to describe complex RNA structures. The ability to describe pseudo-knots — long-range interactions between a loop and another portion of the RNA molecule, as shown in Figure 12-1 — makes PatScan a very powerful program. Table 12-1 lists some of its possibilities.

Table 12-1 RNA-Specific Pattern Rules in PatScan

Rule	*Example*	*Description*
Pairing rule	r1={au,gc} p1=10...12 3...8 r1~p1	Indicates that the p1 pairing only involves au and gc pairs.
Mismatches	p1=10...12 3...8 r1~p1[1,4,2]	The pairing between the two p1 stems can involve at most: 1 mismatch in p1 Vs ~p1 4 unpaired symbols in p1 2 unpaired symbols in ~p1.
Specify patterns	p1=10...12 AUGC~p1	Specifies that the loop MUST contain only AUGC. Note that this pattern must be in uppercase, as opposed to the pairing rules.

With these guidelines, take PatScan on a trial run:

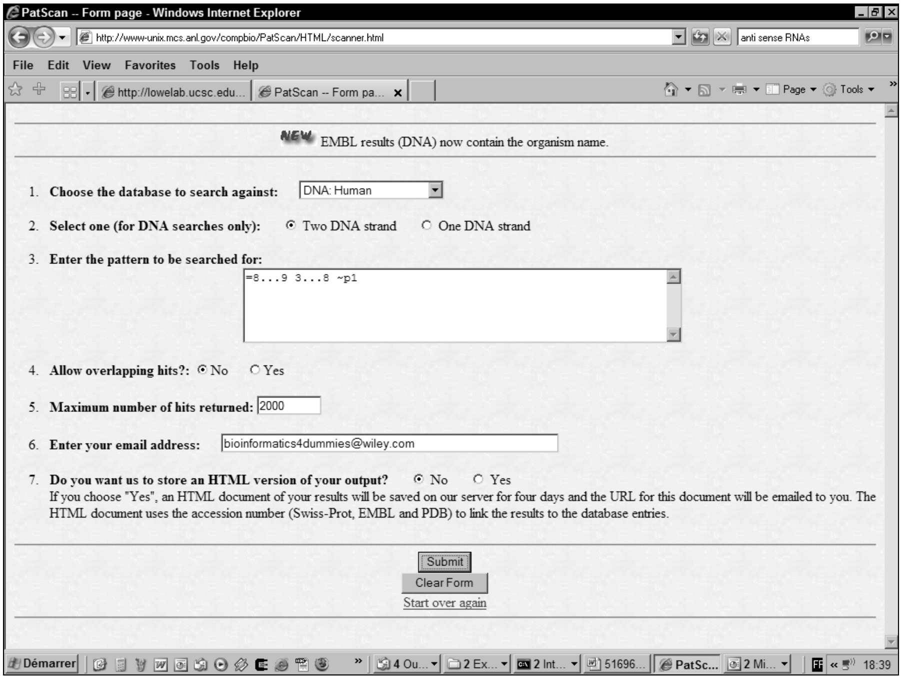
1. **Prepare your PatScan pattern.**
2. **Point your browser to** `www-unix.mcs.anl.gov/compbio/PatScan/HTML/scanner.html`.

   The PatScan server page appears, as shown in Figure 12-8.
3. **Choose the database you want to search from the Database drop-down menu.**

   In our example, we chose DNA: Human.

   The default database on this server is Swiss-Prot. In most cases, this database is *not* appropriate for searching RNA secondary structure patterns.

   You must change this database to a DNA database.


PatScan -- Form page - Windows Internet Explorer
http://www-unix.mcs.anl.gov/compbio/PatScan/HTML/scanner.html
anti sense RNAs
File Edit View Favorites Tools Help
http://lowelab.ucsc.edu...
PatScan -- Form pa...
Page
Tools
Submit
Clear Form
Start over again
Démarrer
4 Ou...
2 Ex...
2 Int...
51696...
PatSc...
2 Mi...
18:39

NEW EMBL results (DNA) now contain the organism name.

1. Choose the database to search against: DNA: Human
2. Select one (for DNA searches only): ● Two DNA strand ○ One DNA strand
3. Enter the pattern to be searched for:

```
=8...9 3...8 ~p1
```

4. Allow overlapping hits?: ● No ○ Yes
5. Maximum number of hits returned: 2000
6. Enter your email address: bioinformatics4dummies@wiley.com
7. Do you want us to store an HTML version of your output? ● No ○ Yes
If you choose "Yes", an HTML document of your results will be saved on our server for four days and the URL for this document will be emailed to you. The HTML document uses the accession number (Swiss-Prot, EMBL and PDB) to link the results to the database entries.

Figure 12-8: Home page of the PatScan server.

4. **Paste or enter your pattern into the Pattern window.**

   We entered p1=8...9 3...8 ~p1.

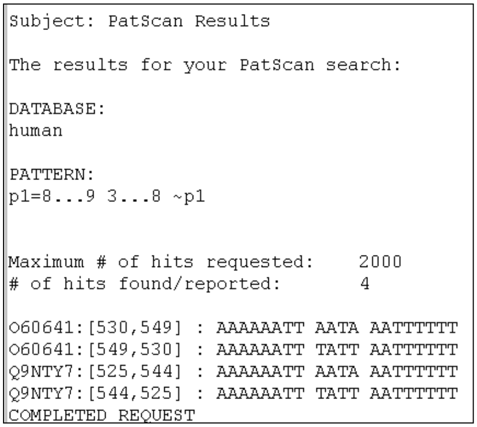
5. **Enter your e-mail address into the appropriate field.**
6. **Click the Submit button.**

   Depending on the size of the database you have selected — and on the length and complexity of your pattern — your search can take up to an hour.

7. **Analyze your results. (See Figure 12-9.)**

   In Figure 12-9, we have used PatScan to search the Human section of GenBank. Each occurrence of a DNA pattern that matches our request appears on a specific line, which indicates the GenBank/EMBL/DDBJ accession number and the coordinates of this match in the sequences.

The fact that a sequence corresponds to your pattern does *not* necessarily mean that it adopts the secondary structure you had in mind! PatScan checks that the hits are compatible with the specified secondary structure, but it doesn't make sure that this structure is the *most* stable one for these hits.



```
Subject: PatScan Results

The results for your PatScan search:

DATABASE:
human

PATTERN:
p1=8...9 3...8 ~p1


Maximum # of hits requested:    2000
# of hits found/reported:       4

O60641:[530,549] : AAAAAATT AATA AATTTTTT
O60641:[549,530] : AAAAAATT TATT AATTTTTT
Q9NTY7:[525,544] : AAAAAATT AATA AATTTTTT
Q9NTY7:[544,525] : AAAAAATT TATT AATTTTTT
COMPLETED REQUEST
```

Figure 12-9: Results of the PatScan server.

Finding the "New" RNAs: miRNAs and siRNAs

One of the most stunning discoveries in biology during recent years was the extent of a phenomenon known as *RNA interference* (*RNAi*) in eukaryotic cells — the possibility for a small, short, non-coding RNA to interfere with a larger, coding RNA — preventing the expression of the corresponding gene. This interference occurs through the binding made possible by complementarity between the small and the long RNA. This is why these small RNAs are sometimes named *antisense RNAs.* (See Chapter 1 and 2 for the basics of nucleic-acid complementarities.) While the possibility of RNA interference had been known for a while in the lab, what stunned everybody was the discovery that nature seems to use it extensively.

Since then, natural short RNAs that can inactivate genes have started popping-up everywhere. They were first discovered in plants and named *silencing RNAs* (*siRNAs*) for their ability to inactivate (silence) genes. These siRNAs are not restricted to plants, and scientists have now found them in most eukaryotic cells. While looking for siRNAs, biologists have started to find a wealth of equally small and unknown RNAs. They have given them the generic name of micro-RNA (miRNA). Nobody really knows what miRNAs really do in the cell, although it's suspected they probably regulate the genes. To make matters worse, not only don't we know exactly what they do, but we also don't know how many of them do it! The reason is that finding miRNAs in a genome is much harder than finding proteins. Nothing makes miRNAs special except for the fact that they are so small. As a consequence, we still don't know how many miRNA genes exist in the human genome. The main distinction between siRNAs and miRNAs is that siRNAs are double-stranded and miRNAs are single-stranded (and usually fold into a stem-loop).

In any case, thanks to the RNA interference phenomenon and the possibility they present of inactivating genes quite specifically, these small antisense RNAs have become one of the favorite tools of molecular biologists and one of the most promising (potential) medical innovations on the horizon — provided somebody ever finds a way to efficiently and accurately send these little guys wherever and whenever they are needed in the human body.

Covering miRNAs, siRNAs, and the phenomenon of RNA interference goes well beyond the *For Dummies* scope — but Table 12-2 gives you enough to start hunting for miRNAs in your favorite genome.

Table 12-2 Hunting Micro RNAs (miRNAs) over the Web

Address	*Description*
`sirna.cgb.ki.se/`	An extensive collection of resources on silencing RNAs.
`itb.biologie.hu-berlin.de/~nebulus/sirna/v2/`	A database of all known human silencing RNAs.
`microrna.sanger.ac.uk/sequences`	The home of miRNAs at the Sanger Center in the UK. Probably one of the most extensive resources on micro-RNAs.
`cbit.snu.ac.kr/~ProMiR2/`	A resource for predicting miRNAs using probabilistic methods.
`pictar.bio.nyu.edu`	Prediction of the potential target of your miRNA on complete genomes.
`bibiserv.techfak.uni-bielefeld.de/rnahybrid/`	A resource for predicting the potential target of your miRNA on a user-provided genomic sequence.
`mirna.imbb.forth.gr/microinspector/`	Runs your genomic sequence against an exhaustive database of miRNAs.

Doing RNA Analysis for Free over the Internet

If you're interested in RNA, everything you need is on the Internet. Make your way through this section to get a better sense of the many excellent resources available online for you.

Studying evolution with ribosomal RNA

To study evolution, you need a gene that exists in all living organisms so you can compare the rate of evolution of these organisms. For many years, biologists have used ribosomal RNA (rRNA) for this purpose. This has become so commonplace that any time you find a new bacterium, the first thing you want to do is sequence its ribosomal RNA so you can see where it fits on the big tree of life.

Several laboratories have dedicated their entire work to helping the scientific community use rRNA sequences. Table 12-3 lists some of the most famous of these resources.

Table 12-3 Ribosomal RNA Resources on the Internet

Address	*Description*
`rdp.cme.msu.edu`	Provides data and services, including the possibility to make online phylogenetic analysis.
`www.psb.ugent.be/rRNA/lsu/`	A European database on the larger of the two ribosomal subunits. It contains predicted structures. It is possible to query the database online. Features lots of online software.
`www.psb.ugent.be/rRNA/ssu/`	The "other" European database, this time dedicated to the small ribosomal subunit.

Finding the small, non-coding RNA you need

Small, non-coding RNAs are poorly represented in major databases. Fortunately, excellent alternative databases make it possible to access the genes you're interested in. Table 12-4 is a partial list of these resources.

Table 12-4 Some Non-Coding RNA Resources

Address	*Description*
`condor.bcm.tmc.edu/smallRNA/smallrna.html`	Dedicated to small non-coding RNAs.

(continued)

Table 12-4 *(continued)*

Address	*Description*
`rna.wustl.edu/tRNAdb/`	Dedicated to tRNAs.
`bighost.area.ba.cnr.it/BIG/UTRHome/`	Dedicated to the untranslated regions of genes.
`www.indiana.edu/~tmrna/`	Dedicated to the recently discovered *tmRNAs* that are both transfer and messenger RNAs. (If you don't yet know what this is, you MUST take a look at this fascinating Web site!)

Generic RNA resources

Well-maintained and usually up to date, these generic RNA resources (see Table 12-5) make it easy for you to locate the server that corresponds to your needs. On these lists of links you can find databases, software, and online servers.

Table 12-5 A List of Generic RNA Resources

Address	*Description*
`bioinfo.lifl.fr/rna/`	A site dedicated to the detection of non-coding RNAs
`www.imb-jena.de/RNA.html`	RNA World, one of the most complete sites currently available
`www.rnabase.org/links/`	Another very complete list of sites

Chapter 13

Building Phylogenetic Trees

In This Chapter

- Finding out what phylogenetic trees can do for you
- Making the right multiple sequence alignment for the right tree
- Estimating distances between your sequences
- Building a Neighbor Joining Tree
- Evaluating the quality of a tree by bootstrapping
- Finding tree resources on the Internet

> *Nothing in biology makes sense except in the light of evolution.*
>
> — Theodosius Dobzhansky (1973)

The purpose of this chapter is to show you how you can use a computer and a few online services to run a *phylogenetic analysis,* the scientific procedure that lets you reconstruct the evolutionary history of a group of organisms or sequences. If you know what phylogenetic trees are all about — but you don't have a clue how to generate one with a computer — this chapter is for you.

Here we show you how you can select sequences, turn them into a multiple sequence alignment, and then turn this alignment into a tree (well, okay, a diagram that *looks* like a tree). We show you three methods:

- **ClustalW:** This is an easy method that you must use as a black box. You cannot really control ClustalW, but it is reassuring to know that it uses Neighbor Joining, a fairly well-established tree method.
- **Phylip:** A more sophisticated method that enables you to control every parameter that plays a part in the making of your tree.
- **phyML:** A very accurate reconstruction method that uses maximum likelihood, a method that specialists often consider to be the most accurate. As its name implies, Maximum Likelihood methods try to reconstruct the tree that more likely to correspond to your alignment.

Whatever choice you make, bear in mind that a high-quality multiple sequence alignment is usually the actual limiting factor. (If you're not sure what a multiple sequence alignment is, be sure to check out Chapter 9.)

Finding Out What Phylogenetic Trees Can Do for You

The purpose of *phylogeny* is to reconstruct the history of life and explain the present diversity of living creatures. This can be represented as a huge genealogic tree (the tree of life). The underlying principle of phylogeny is to try to group living creatures according to their level of similarity. In this context, we assume that the more similar two species are (such as human and ape), the closer they are to their common ancestor. Phylogeny is not a new subject, and whether you trace back the birth of modern biology to Darwin and *The Origin of Species* or to Aristotle and his notion of categories, you can't escape this daunting fact: Biology is very much about classifying — and the best means of classification we have is phylogeny.

Phylogenetics is a special kind of phylogeny that relies on the comparison of equivalent genes coming from several species for reconstructing the genealogic tree of these species and finding out who is the closest relative of whom in the family. If necessary, you can also apply phylogenetic methods to the various genes of a gene family to reconstruct the history of the gene family by the same means. (Take note that these trees make sense only if you believe in evolution!)

The molecular stories you can uncover with phylogenetics are incredibly rich. In fact, we find it difficult *not* to see a parallel between the destiny of famous families (such as the Medicis, the Borgias, or the Kennedys) and the fates of gene families. When we unravel the chain of events that makes the story of a protein family, we find tales of mutations and deletions, duplications or speciation, loss and gain of function, inactivation, and all the other traumatic events that shaped the world as it is today. Nothing is ever taken for granted while life evolves!

Phylogenetics is here to let you discover all this. Phylogenetics is not a technique, nor even a discipline; phylogenetics is a science — and a major one. The mere task of laying out the general ideas of phylogenetics is way beyond the scope of this chapter, so if this subject is new to you, we urge you to consult some of the excellent textbooks available on this subject.

In the context of bioinformatic analysis, there are three major reasons why you may want to use phylogenetics:

- **Determining the closest relatives of the organism that you're interested in:** For instance, if you're studying a new bacterium, you can sequence its ribosomal RNA and place it on a phylogenetic tree computed with all known ribosomal RNAs. This can give you a fairly good idea of who this bacterium really is.
- **Discovering the function of a gene:** If you're studying a gene, you can use phylogenetic trees to be sure that the gene you're interested in is orthologous (more about that in a minute) to another well-characterized gene in another species.
- **Retracing the origin of a gene:** Most genes within a genome travel together through evolutionary time. However, from time to time, individual genes may jump from one species to another — for instance, piggybacking a virus infection. Phylogenetic trees are a great way to reveal such events, which are called *horizontal* (or *lateral*) *transfers.*

Two genes are *orthologous* if they come from two different organisms, derive from a common ancestor gene, and have been separated only by speciation events (as opposed to gene duplications). In theory, two genes that are orthologous often have the same exact function (have similar roles) in the two different organisms they come from. On the other hand, when the two genes arose from gene duplication, we say that they are *paralogous.* Two genes that are paralogous belong to the same family but are more likely to have different functions than they would if they were orthologous.

To give you an image, if the joystick of a Boeing 747 were a gene, it would be orthologous with the steering wheel of your car. In contrast, the left front wheel of your car is paralogous with the right front wheel.

Preparing Your Phylogenetic Data

Biologists have been dealing with phylogeny for many decades now, hence the large number of different techniques available. Things have become a bit simpler since people started doing *phylogenetics* (phylogeny applied to gene data). Nonetheless, there are still many schools of thought around — and ancient fights are still raging in the world of phylogeny. Be aware that when you choose a method, you're also choosing sides! In this chapter, we try to focus on the methods that have attracted the smallest amount of controversy over the past few years.

The one thing upon which everybody agrees is that data quality is critical. If you have very good data that's suitable for phylogenetic analysis, any reasonable method will lead to a reasonable tree. On the other hand, if your data isn't good, no method — however good — can rescue the story your data tells. A brutal summary of this observation exists: *Garbage in, garbage out.* This is true of most data-analysis methods.

Phylogenetics uses genes to reconstruct evolutionary history. This type of data is much easier to interpret than the *morphology data* — derived from an analysis of an organism's form and structure — that biologists used to rely on. It's possible to compare gene sequences in a much more objective fashion than it is to compare morphological characteristics.

When you do phylogenetics, good data means *a highly accurate multiple sequence alignment that contains properly chosen sequences.* In the first part of this section, we show you how you can prepare this kind of data.

Choosing the right sequences for the right tree

When you build a phylogenetic tree, you make the assumption that the sequences you are comparing have a common ancestor. If your sequences are similar enough, this is a reasonable hypothesis. You want to use your tree to reconstitute the history of these sequences to find out who is the closest relative to whom.

Of course, this vision implies that evolution is always divergent — thus if two sequences are similar, we rule out the possibility that this similarity may result from convergent evolution. In the vast majority of cases, this rule makes sense. At the molecular level, convergent evolution does exist — but it seems to be the exception rather than the rule.

Using DNA or protein sequences

To establish the relationship between two sequences, you want to measure the time that separates their divergence from their common ancestor. How long ago did they part? To answer this question, compare your sequences and measure a distance or establish an evolutionary scenario. Should you do this on the protein or on the DNA sequence? We have a complex answer to this seemingly simple question:

- **If your DNA sequences are *more* than 70 percent identical:**

  You can make a DNA multiple sequence alignment. If your sequences are coding for proteins, however, this is not recommended.

The problem with DNA sequence alignments is that the substitution matrices are not very good — and the alignments take more time to compute because of the sequence lengths.

- **If your DNA sequences are *less* than 70 percent identical:**

  If your sequences code for proteins, it is much safer to translate them into proteins and to build the multiple sequence alignment with the proteins.

  If your sequences are too similar at the protein level, you can thread the DNA sequences back onto the protein alignment. Use pal2nal for that purpose — like this:

  ```
  coot.embl.de/pal2nal
  ```

  If you do not have the DNA sequence that corresponds to your proteins, you can use Protogene to automatically fetch the bona-fide DNA sequences of your proteins. Protogen is a Tcoffee tool, available at `www.tcoffee.org`

  After you've completed the DNA to protein threading, you have two possibilities:

  - **If most synonymous mutation sites are different:** This is a typical case of saturation. In this case, the distance measurements you make at the protein level are more accurate.
  - **If synonymous sites are not saturated:** Distance measurements made at the DNA level are more accurate in theory.

  In practice, unless your sequences are almost identical, it is easier to keep working at the protein level. This may not be as accurate as working with DNA sequences, but, in most cases, you can expect the results to be reasonably good.

White T-shirts, spaghetti, evolution, and phylogeny

When we think of evolution, we tend to think of natural selection, survival of the fittest, struggle for life, and all these testosterone-loaded concepts. In molecular terms, gurus name this *neo-Darwinism.* This popular (though ancient) theory states that the only reason mutations stay in a population is because they are helpful and increase the fitness of the individuals that carry them. But this theory doesn't account very well for synonymous mutations or other harmless DNA alterations.

In the 1960s, the increased availability of molecular data led to Motoo Kimura's elaboration of neutralism. *Neutralism,* also known as the neutral theory, challenges neo-Darwinism by stating that most mutations are either lethal or neutral. Everybody agrees that lethal mutations are disposed of right away — but for neutralists, the possibility of a new non-lethal mutations spreading (or not spreading) in a population depends mostly on chance. Even if these mutations are advantageous, chance is the driving

(continued)

(continued)

force. Keep in mind that neutral mutations aren't necessarily silent at the protein level; some such mutations change a protein slightly without altering its function. Of course, after these harmless-but-useless mutations are fixed, they can also find a niche — and give some advantage to the individuals who carry them — later on. Neutralism is a very powerful theory because it accounts well for the way in which species build their adaptation potential before they need it.

If you're doing phylogeny, these neutral mutations are really the ones you're interested in. You do not want to see adaptation; you want to measure evolutionary distances and the best measure of evolutionary distance you have is counting the number of random and quasi-neutral events that occurred on your sequences since they separated from their ancestor. You can think of these mutations like each tick of a clock that has been working, one mutation at a time, for a million years.

To give you an image, imagine that this morning, like every day, you took a white T-shirt and duplicated it for your two kids (Dave and Benny), and off they went. It was a long day; Dave and Benny both ate their spaghetti bolognese, both proving especially adept at emulating Michelangelo with the verve and passion with which they sketched out — in brilliant red sauce — a few Last Judgments on their white T-shirts. Now suppose that each drop of colored-something landing on the immaculate shirts is a mutation — and the number of spots indicates how far the day has progressed. If you want to know how long Dave and Benny have been wearing these shirts, you can simply count the spots. This is true neutralism, and this is the ideal situation when you compare sequences.

Imagine now that there is a bit of selection going on: Dave and Benny may be very anxious about your reaction when they come back (okay, okay, probably unlikely, but bear with us for the sake of argument), and they carefully clean off every bit of color that lands on their shirts — except for those you can't see or those that look "coooool." This would be the equivalent of selected mutations. When this goes on, you can no longer tell how long the shirt has been worn.

Non-selected mutations (neutral) are much better for evaluating distances because they accumulate smoothly. In coding DNA, the synonymous mutations are the more typical non-selected mutations. This is why DNA is better for measuring distances than protein sequences whose evolution is much more constrained. (These protein guys have work to do!)

Imagine now that the kids have had a second or third go at the spaghetti bolognese. Drops of sauce have landed on previous spots, and the shirts are now completely saturated with tomato sauce. What can you tell from this? How many spaghetti meals have Benny and Dave had? One? Two? Five? It's impossible to say!

This analogy is typical of how our evolutionary time measure becomes saturated — when there have been one or more mutations on every neutral site of our DNA sequence. Determining how long this has been going on becomes impossible. Of course, the closer you are to saturation, the more imprecise your measurements are going to be.

One more thing before we rush to wash these filthy T-shirts: What if, on a certain day, Benny and Dave have not been doing the same thing? One may come home sparkling clean and the other muddy. Will you assume that they have spent their days in different time dimensions, or would you rather conclude that their shirts have evolved at different paces (or places)? Different paces sound more likely — although, if you ask the kids, you may get a more confusing answer. Such differences also occur with genes — and although we like to think that there is a universal molecular clock, we know it tends to tick at a different pace in every gene.

Choosing sequences to make either a gene tree or a species tree

Homologous genes are genes that derive from a common ancestor. They can have three types of relationships:

- **Orthologs:** They're only separated by *speciation* — is the phenomenon during which a common ancestor gives birth to two subgroups that slowly drift away from their common genetic makeup to become distinct species. Assuming that the genomes are not rearranged in the two new species, two genes are orthologous when they correspond to the same ancestral gene in the ancestral genome. Biologists usually expect orthologs to have similar functions and structure. In Figure 13-1, A1 and A2 are orthologs, and so are B1 and B2.
- **Paralogs:** Paralogs are homologues separated by a duplication event, meaning that within a genome, a gene was duplicated. One of the duplicates may have kept the original function while the other duplicate could have acquired a new function. You can expect paralogs to have different but related functions. For instance, A1 and B1 are paralogs in Figure 13-1.
- **Xenologs:** *Xeno* is a Greek word that means "foreigner." Xenologs result from a *lateral transfer* between two organisms — a direct DNA transfer between two species. This means that one of the species contains a gene that does not have the same history as the genome in which it is inserted. A typical case of lateral transfer (or xenologs) is the acquisition of the isoleucyl-tRNA sytnthase from their host by several bacteria. The isoleucyl-tRNA sytnthase is a protein involved in the synthesis of other proteins, and its acquisition by bacteria seems to help them becoming antibiotic resistant. When this happens, the newly acquired isoleucyl-tRNA sytnthase is a xenolog of the other tRNA synthases contained in the bacteria.

When you select a group of homologous genes to make a phylogenetic tree, you always make what biologists call a *gene tree.* It is a tree that tells the story of the genes it contains.

If you select all the *paralogous* members of a large human gene family, your gene tree tells the story of this gene family only. You can only use it to reconstruct the chain of duplications that led from one single ancestral gene to the current situation.

If you select a group of genes that are all *orthologous* from different species, the gene tree you get looks very much like a species tree — which lets you reconstruct the speciations that occurred while the species you're looking at (or their ancestors) were diverging. The best example of this type of gene tree is the ribosomal RNA phylogenetic tree that biologists use to reconstruct the big tree of life. Ribosomal RNA genes exist in every species and are clearly orthologous between species.

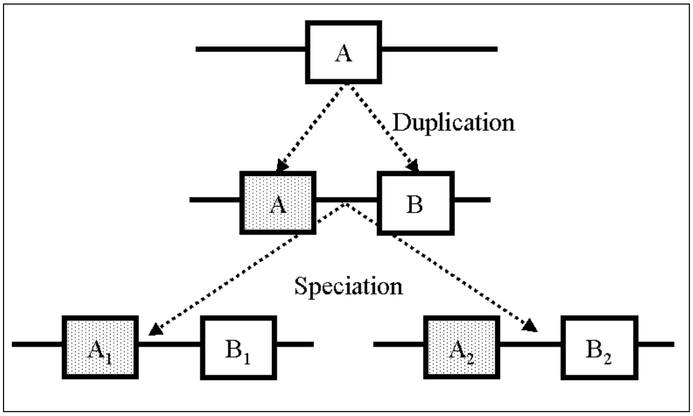


Figure 13-1: Orthology and paralogy.

You can do things the other way round, of course, and use the tree to ask whether your sequences can be called orthologs or paralogs. This is a valid strategy if you know the true species tree.

Two genes that are homologous and come from two different species are not necessarily orthologous. They can also be paralogous — as A1 and B2 are in Figure 13-1. Mixing orthologs and paralogs in a phylogenetic tree is a major source of trouble. Unfortunately, avoiding it is difficult.

In theory, there's no simple solution to the problem of establishing orthology — and demonstrating it is virtually impossible. However, in practice, authors have been using an empirical criterion that works reasonably well: the reciprocal BLAST best match. Here's how it's done:

1. **Choose a sequenceA from genomeA.**
2. **Carry out a BLAST search comparing your sequenceA and every sequence in another complete genomeB. (For more on BLAST searches, see Chapter 7.)**

   The BLAST search returns sequenceB as a top hit.
3. **Carry out a BLAST search comparing sequenceB and every sequence in genomeA.**

   If this search returns sequenceA as a top hit, you can treat sequenceA and sequenceB as orthologs from genomeA and genomeB.

   This result doesn't demonstrate that sequenceA and sequenceB actually *are* orthologous, but it does mean you won't be doing something very wrong if you consider them so.

WARNING!

This strategy makes sense only when your two genomes are *complete* and have roughly the same proteome size.

Eugene Koonin has assembled a high-quality collection of orthologous genes at NCBI. To do so, he used similar (but slightly more sophisticated) criteria. This collection is known as the COG (Clusters of Orthologous Groups), and you can find it online at `www.ncbi.nlm.nih.gov/COG/`. Other collections of homologous genes include HOGENOM (former HOBACGEN) and HOVERGEN developed by the Pôle Bioinformatique Lyonnais (`http://pbil.univ-lyon1.fr/`).

Creating the perfect set

"Perfect" is meant to be ironic here, but we've listed in Table 13-1 a few points you should take into account when assembling your sequence set.

Table 13-1 Preparing a Set of Sequences for Making a Phylogenetic Tree

Problem	*Reason and Solution*
Avoid sequence fragments	Incomplete sequences make multiple sequence alignments and tree reconstruction methods very sick! You want to avoid these at all costs — or you at least want to use the same fragment for all the sequences.
Avoid Xenologs	Unless your purpose is to study them, avoid including genes that result from lateral transfer.
Avoid recombinant sequences	Some proteins result from the combination of several proteins. This is especially common in viruses. In terms of phylogeny, such proteins have two ancestors rather than one — and standard tree methods are not equipped to represent this kind of relationship.
Avoid large complex families	Very large families that contain various domains and repeats can be very tricky to analyze. The ABC transporter is a good example of this kind of troubled family. If you can, stay away from these protein gangs — or try to work on smaller, more uniform subsets.

(continued)

Table 13-1 *(continued)*

Problem	*Reason and Solution*
Try to make a small set	Phylogenetic reconstruction is one of the only cases where you're allowed to align very similar sequences — as many as you want. Unfortunately, it isn't because you need it that you can do it. If you've been through Chapter 9, you know this is not a good thing because most methods find it hard to deal with large datasets. Analyzing large sets means that you may not be able to use the most powerful methods, or that you will need to install these programs on your computer.
Add an outgroup to your dataset	An *outgroup* is a sequence that you know has diverged long ago from the rest of the set. For instance, horse hemoglobin would be an outgroup if you're analyzing primate hemoglobin. This outgroup helps you by adding a root to your tree that symbolizes the first common ancestor of all your sequences. The choice of an outgroup relies mostly on biological criteria, yet you must also make sure that this sequence is similar enough to the rest of the set so your multiple-sequence-alignment method can align it (see Chapter 9).

REMEMBER

Living fossils do not exist! Unless you are using fossil DNA sequences, the sequences you are comparing all have the same age. Even if one of the species has retained more similarity with the common ancestor, its sequences have accumulated as many neutral mutations as the most sophisticated member of the family.

Preparing your multiple sequence alignment

Many phylogenetic tree-reconstruction methods require a special table called *distance matrix.* This table contains the distances (or counts the number of evolutionary events) that separate each pair of sequences in your dataset.

In theory, you can build this matrix by comparing all your sequences two by two and then filling up the distance matrix. In practice, that doesn't work — the pairwise comparison methods aren't accurate enough. (For more on pairwise comparisons, see Chapter 8.) The only way to accurately measure the distances between two sequences is a two-step process:

1. **Make a multiple sequence alignment that includes the two sequences.**
2. **Measure the distances on the sequences' pairwise projection.**

   The *pairwise projection* is a pairwise alignment that you *extract* from the multiple sequence alignment. It benefits from the information contained in ALL the other sequences.

The quality of your multiple sequence alignment is the real limiting factor when you make a tree; there is no way you can make a good tree with a bad alignment. (If you aren't sure what a multiple sequence alignment *really* is, make your way through Chapter 9.)

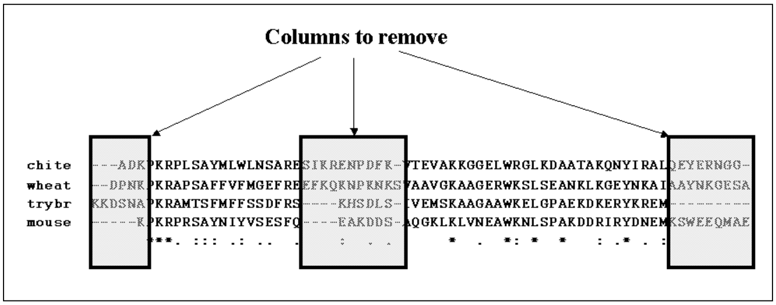
Computing your multiple sequence alignment

To compute your multiple sequence alignment, you can use ANY of the methods we introduce in Chapter 9. These online servers include

- **ClustalW:** `www.ebi.ac.uk/clustalw`
- **MUSCLE:** `phylogenomics.berkeley.edu/cgi-bin/muscle/input_muscle.py`
- **Tcoffee:** `www.tcoffee.org`

Making sure you have the right multiple sequence alignment

Before using your multiple sequence alignment for building a tree, you want to make sure that it's as accurate as possible. In this section, we give you a few criteria — and a list of things to do — to improve the suitability of your multiple alignment. (Figure 13-2 shows you the kinds of regions our handy steps list below targets for removal from your alignment.)



```
chite  ---ADKPKRPLSAYMLWLNSARESIKRENPDFK-VTEVAKKGGELWRGLKDAATAKQNYIRALQEYERNGG-
wheat  --DPNKPKRAPSAFFVFMGEFREEFKQKNPKNKSVAAVGKAAGERWKSLSEANKLKGEYNKAIAAYNKGESA
trybr  KKDSNAPKRAMTSFMFFSSDFRS----KHSDLS-IVEMSKAAGAAWKELGPAEKDKERYKREM---------
mouse  -----KPKRPRSAYNIYVSESFQ----EAKDDS-AQGKLKLVNEAWKNLSPAKDDRIRYDNEMKSWEEQMAE
             ***. ::: .: ..  .    :  . .      *  .  *: *  *   : .* . :
```

Figure 13-2: Preparing a multiple sequence alignment.

To ensure that your multiple alignment is both accurate and suitable, follow these steps:

1. **Make sure there are as many gap-free columns as possible.**

   Gaps cause trouble for most phylogeny reconstruction methods. Some of these methods, such as ClustalW, can use the *complete-deletion* techniques and ignore every column that contains a gap.

2. **Remove the extremities of your multiple alignment.**

   The N-terminus and the C-terminus tend to be poorly conserved — and therefore poorly aligned. You can safely remove them (refer to Figure 13-2).

3. **Remove the gap-rich regions of your alignment.**

   Internal, gap-rich regions in a multiple sequence alignment often correspond to loops. Even if your program returns an alignment, it does not mean that this alignment is meaningful.

4. **Be sure to keep the most informative blocks.**

   The ideal multiple alignment for building a tree would be a high-quality alignment of sequences with a low level of identity, so that each position contains a trace of the family history. Here are some pointers:

   - In Chapter 9, we show you what a good block looks like in a multiple sequence alignment. It's typically 20 to 30 amino acids long, and contains a few conserved positions. Such blocks are ideal for producing high-quality trees.
   - If you want, you can use the Tcoffee server to evaluate your multiple sequence alignment and remove columns that are unlikely to be correctly aligned. (See Chapter 9 for more on how to do this.)
   - The best way to edit your multiple alignment is to use Jalview, the multiple alignment editor that we introduce in Chapter 10.

   If you want a quick-and-dirty solution, here's a way to use Microsoft Word to remove blocks very efficiently.

   a. While pressing the Alt key on your keyboard, use the mouse to select entire columns in your alignment.

   b. When you've selected everything you want to remove, press the Delete key to remove the selected block.

Don't even *try* to do anything else to your sequences and alignments while using a word processor! If you need to edit your sequences, use a proper alignment editor like Jalview (Chapter 10):

Building the Kind of Tree You Need

There are three major groups of methods you can use to build a tree: distance methods, parsimony methods, and likelihood methods. None of these methods is almighty. They all have their particular weaknesses and strengths.

This chapter shows you how to make distance-based and maximum-likelihood trees. Distance-based trees aren't necessarily the most accurate, but they are generally applicable and easy to set up. You can apply distance-based tree-reconstruction methods on most situations; they tend to generate very reasonable results. Maximum Parsimony and Maximum Likelihood methods are (in theory) the most accurate — but they take more time to run. Experts tend to give a slight hedge to Maximum Likelihood, but the debate is still largely open.

When you decide to use a distance-based method to compute your phylogenetic tree, consider these four major ingredients:

- **Your multiple alignment:** Which method? Which sequences?
- **The distance measure:** Which substitution matrix?
- **The tree-reconstruction algorithm:** UPGMA, NJ, Fitch, or Kitsch?
- **The type of tree you want to display:** A tree without distances (called a *cladogram*) or a tree displaying distances (called a *phenogram*)? A rooted or an unrooted tree?

In this section, we show you how you can use the Internet to compute a tree and to display it in any format you need.

Computing your tree

Combining the right ingredients for making your tree is much like cooking. There is really no such thing as *the best recipe ever;* the best tree you can produce is mostly a matter of taste and the mood of the day.

As when it comes to food, there are two very different ways to produce trees. The first one uses ClustalW; it's quick, hassle-free, and somehow very similar to (good) fast food. We show you in a steps list how you can simply cut and paste your sequences to obtain a tree while avoiding embarrassing questions.

The second step list is for those of you out there who love to buy fresh vegetables on the market and make your own salad dressing — sticklers for detail, if you catch our drift. With these steps, you can control every ingredient that goes into the making of your tree. Phylip is the name of this package that you can use to do state-of-the art phylogeny. This is serious stuff; world-renowned phylogeny specialists use it to do their work. Despite this intimidating introduction, Phylip is surprisingly easy to use.

Producing an (almost) instant phylogenetic tree with ClustalW

If you need a tree and you have firmly decided that you do not want to know *anything* about this barbaric kind of gardening, ClustalW is definitely for you. ClustalW is, above all, a multiple-sequence-alignment package, but some of the ClustalW servers let you use this program to produce phylogenetic trees.

On the EBI ClustalW server — and on most ClustalW Web servers — if you feed ClustalW some unaligned sequences, the program returns a tree. This tree you get, however, is **NOT** a phylogenetic tree; it is a guide tree that ClustalW uses to assemble the multiple sequence alignment. In general, any tree with a `.dnd` extension returned by ClustalW is NOT a phylogenetic tree.

You CANNOT use this tree in place of a phylogenetic tree. Doing so would be a mistake (as in: incorrect results and a false sense of security).

In this section, we show you how you can use the EBI ClustalW server to produce a phylogenetic tree. We assume you have already generated your multiple sequence alignment; if you have not, you can download a dummy alignment from `www.tcoffee.org/dummy_aln2.html`.

Our dummy alignment contains homologous fragments of the gamma fibrinogen. They have been sequenced in several species for comparative studies. We can use fragments rather than entire sequences because these fragments are all homologous on their entire length. We have renamed the fragments with the species names, since trees with species names are much easier to read — unless, of course, you have a formidable memory,

O12957 (Sheep), O02672 (Moose), O02683 (Giraffe)

O02690 (Chevrotain), O02681 (Beluga) O02687 (Sperm_whale)

O02673 (Rorqual), O02688 (Pig), O12959 (Peccary)

O02677 (Dromedary), O02689 (Tapir), O02682 (Horse)

O02676 (Hyena), O02680 (Coyote), O12954 (Hippopotamus)

Of course, you could use any set of sequences aligned with any of the servers we described in Chapter 9.

1. **Point your browser to** `www.ebi.ac.uk/clustalw/`.

2. **Paste your multiple alignment into the Sequence window, as shown in Figure 13-3.**

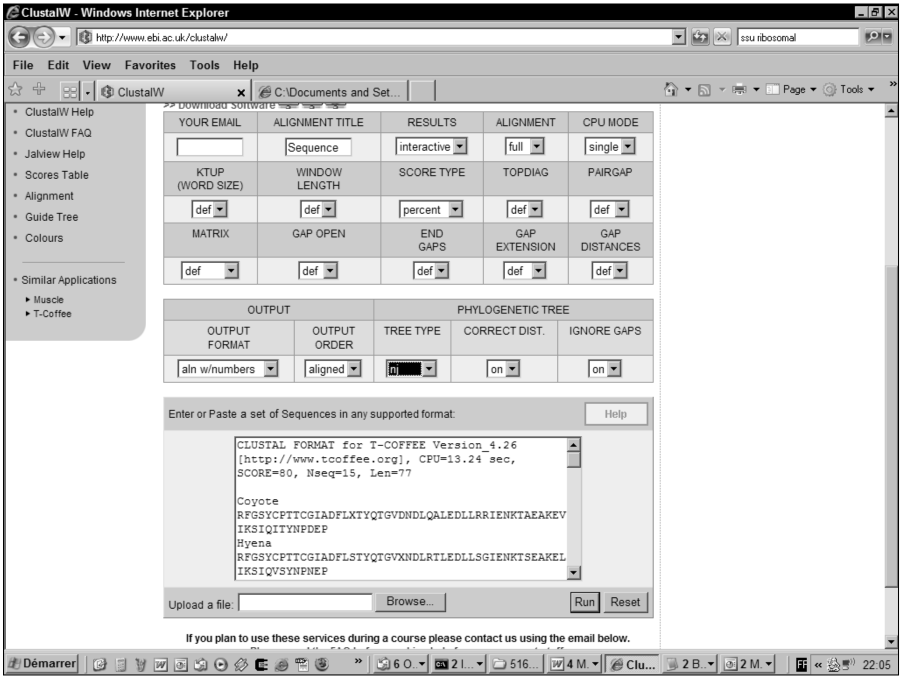
   You can enter your alignment in any format you like, including ALN (the ClustalW format), MSF, or FASTA. If you use an ALN alignment, make sure you include the first line that reads "CLUSTAL W ...(etc.)"

3. **Choose NJ from the Tree Type drop-down menu in the Phylogenetic Tree section, as shown in Figure 13-4.**

   This menu enables you to choose any of several different methods for computing your tree. Specialists generally agree that NJ, the Neighbor Joining method, is the best one for the task at hand.

WARNING!

If you set the Phylogenetic Tree: Tree Type drop-down menu to anything other than None, you must provide ClustalW with aligned sequences. "Why?" you might ask? Well, when you choose anything but None from the drop-down menu, ClustalW *stops computing alignments* and assumes that you have provided it with aligned sequences. If you don't do as it assumes — and simply give the server unaligned FASTA sequences — it assumes the sequences *are* aligned and computes a tree. Of course (unless your sequences can be aligned without gaps), this tree will be wrong!



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	nj	on	on

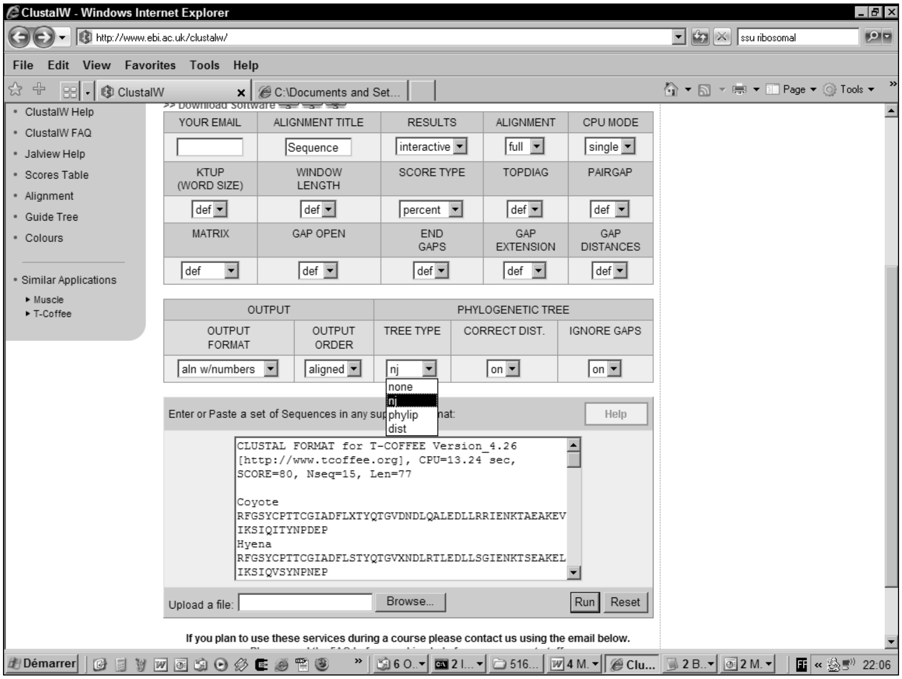
Enter or Paste a set of Sequences in any supported format: Help

```
CLUSTAL FORMAT for T-COFFEE Version_4.26
[http://www.tcoffee.org], CPU=13.24 sec,
SCORE=80, Nseq=15, Len=77

Coyote
RFGSYCPTTCGIADFLXTYQTGVDNDLQALEDLLRRIENKTAEAKEV
IKSIQITYNPDEP
Hyena
RFGSYCPTTCGIADFLSTYQTGVXNDLRTLEDLLSGIENKTSEAKEL
IKSIQVSYNPNEP
```

Upload a file: Browse... Run Reset

If you plan to use these services during a course please contact us using the email below.

Figure 13-3: Computing a phylogenetic tree with ClustalW.



YOUR EMAIL	ALIGNMENT TITLE	RESULTS	ALIGNMENT	CPU MODE
	Sequence	interactive	full	single
KTUP (WORD SIZE)	WINDOW LENGTH	SCORE TYPE	TOPDIAG	PAIRGAP
def	def	percent	def	def
MATRIX	GAP OPEN	END GAPS	GAP EXTENSION	GAP DISTANCES
def	def	def	def	def

OUTPUT		PHYLOGENETIC TREE		
OUTPUT FORMAT	OUTPUT ORDER	TREE TYPE	CORRECT DIST.	IGNORE GAPS
aln w/numbers	aligned	nj	on	on

Enter or Paste a set of Sequences in any sup nat:

```
CLUSTAL FORMAT for T-COFFEE Version_4.26
[http://www.tcoffee.org], CPU=13.24 sec,
SCORE=80, Nseq=15, Len=77

Coyote
RFGSYCPTTCGIADFLXTYQTGVDNDLQALEDLLRRIENKTAEAKEV
IKSIQITYNPDEP
Hyena
RFGSYCPTTCGIADFLSTYQTGVXNDLRTLEDLLSGIENKTSEAKEL
IKSIQVSYNPNEP
```

Upload a file:

If you plan to use these services during a course please contact us using the email below.

Figure 13-4: The Phylogenetic Tree: Tree Type drop-down menu in ClustalW.

4. **Choose On from the Correct Dist. drop-down menu (located directly to the right of the Phylogenetic Tree: Tree Type drop-down menu).**

   This option makes it possible to correct for multiple substitutions. It tends to elongate the branches of your tree; it's mostly useful for distantly related sequences.

5. **Choose On from the Ignore Gaps drop-down menu.**

   This causes ClustalW to ignore every column of your multiple alignment that contains gaps. Turning this option on ensures that *all* sequences are compared on the same number of residues. Keep in mind that phylogeny gurus usually prefer to ignore positions (columns) that contain gaps.

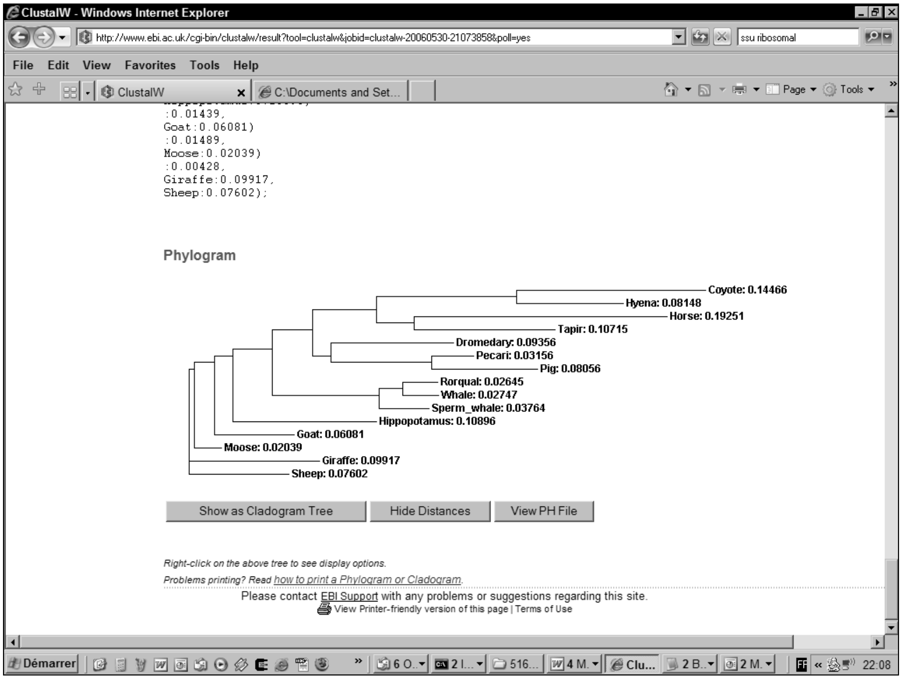
6. **Click the Run button.**

   Wait until ClustalW returns your NJ tree, as shown in Figure 13-5.

   This tree is much more accurate than a guide tree, and this one clearly shows the genetic relationship between the hippopotamus and the whale, as postulated by Higgins and Grauer a few years ago.

7. **Save the graphical representation of your tree.**

   The easiest way to save your tree is to make a screen capture with the print-screen (Prt Sc) key on your keyboard. You can then cut and paste this image into your favorite application (PowerPoint, Paint. etc.).



```
:0.01439,
Goat:0.06081)
:0.01489,
Moose:0.02039)
:0.00428,
Giraffe:0.09917,
Sheep:0.07602);
```

Right-click on the above tree to see display options.

Problems printing? Read how to print a Phylogram or Cladogram.

Please contact EBI Support with any problems or suggestions regarding this site.
View Printer-friendly version of this page | Terms of Use

Figure 13-5: An NJ tree produced by ClustalW.

8. **Scroll to the bottom of the page and click on View PH File.**

9. **Use the File➪Save As option of your browser to save a text version of your tree.**

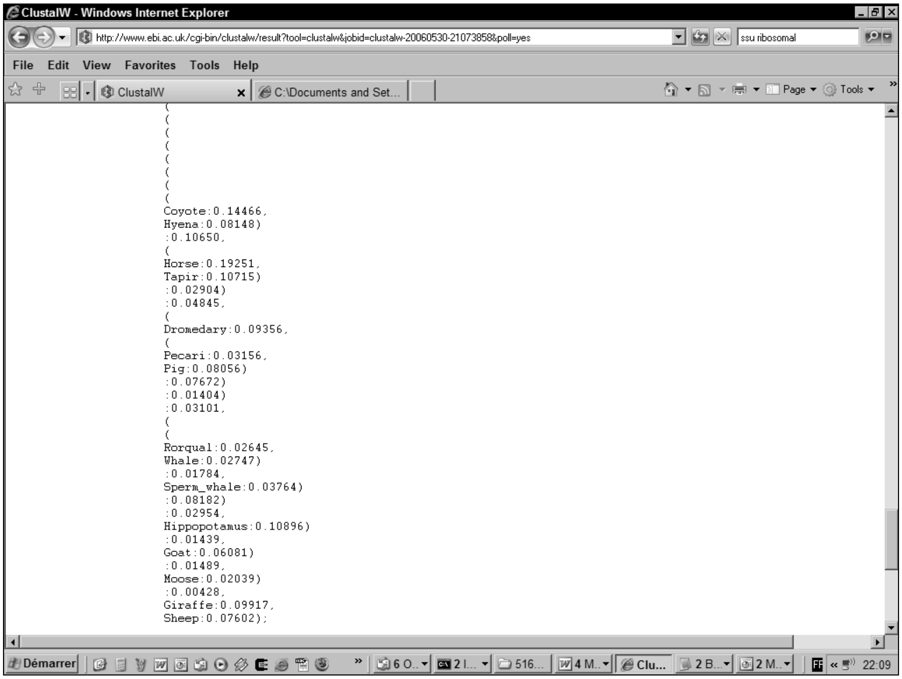
   ClustalW presents your tree in the Newick text format — shown in Figure 13-6 — famous for its many parentheses. When you save the file, you're saving this Newick-formatted version of your tree.

   The Newick format is the most standard way to store phylogenetic trees, and most phylogenetic packages can use this format. If you want to keep only one file related to your tree, keep this one!

Making a tree with Phylip

When it comes to phylogenetic trees, Phylip is a household name. Along with PAUP, it is one of the most widely used resources for computing highly accurate phylogenetic trees.

One of the nice things with Phylip is that it makes it easy for you to *bootstrap* your tree (check its reliability; see the sidebar, "What is bootstrapping?"). What that does for you is make sure that the topology you're looking at is well supported by your multiple sequence alignment.



ClustalW - Windows Internet Explorer
http://www.ebi.ac.uk/cgi-bin/clustalw/result?tool=clustalw&jobid=clustalw-20060530-21073858&poll=yes
ssu ribosomal
File Edit View Favorites Tools Help
ClustalW
C:\Documents and Set...
Page
Tools
Démarrer
22:09

```
(
(
(
(
(
(
(
(
Coyote:0.14466,
Hyena:0.08148)
:0.10650,
(
Horse:0.19251,
Tapir:0.10715)
:0.02904)
:0.04845,
(
Dromedary:0.09356,
(
Pecari:0.03156,
Pig:0.08056)
:0.07672)
:0.01404)
:0.03101,
(
(
Rorqual:0.02645,
Whale:0.02747)
:0.01784,
Sperm_whale:0.03764)
:0.08182)
:0.02954,
Hippopotamus:0.10896)
:0.01439,
Goat:0.06081)
:0.01489,
Moose:0.02039)
:0.00428,
Giraffe:0.09917,
Sheep:0.07602);
```

Figure 13-6: The Newick tree format.

If you decide to become a specialist, Phylip has all you need; you can use it to build almost any kind of tree with any method you like. To demonstrate, here's a set of steps that shows how to produce a phylogenetic tree with bootstrap information:

Before going onto this server, you must have a high-quality multiple sequence alignment ready. For the purpose of these steps, we have prepared one for you, available from `www.tcoffee.org/dummy_aln2.html`.

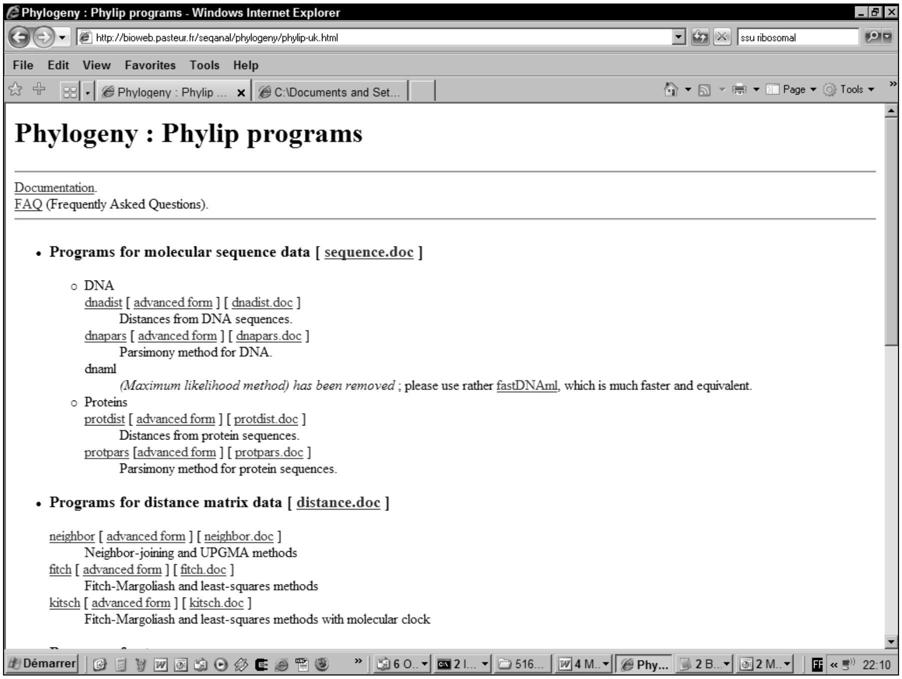
1. **Point your browser to** `bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`**.**

   The Phylogeny: Phylip Programs page appears, as shown in Figure 13-7. This page shows all the Phylip programs installed on this server. You can produce a phylogenetic tree by using these programs one after the other. The only trick is to use them in the right order.

2. **Under the Proteins heading, right next to the protdist link, click the advanced form link.**

   The Phylip: Protdist page appears. Here's what you need to know:

   - To make a distance-based tree, you first need to generate a distance matrix that contains the pairwise distances between sequences, as measured on the multiple sequence alignment.


Phylogeny : Phylip programs - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html
ssu ribosomal
File Edit View Favorites Tools Help
Phylogeny : Phylip ...
C:\Documents and Set...
Page
Tools
Démarrer
22:10

Phylogeny : Phylip programs

Documentation.
FAQ (Frequently Asked Questions).

- **Programs for molecular sequence data [sequence.doc]**
  - DNA
    dnadist [advanced form] [dnadist.doc]
    Distances from DNA sequences.
    dnapars [advanced form] [dnapars.doc]
    Parsimony method for DNA.
    dnaml
    (Maximum likelihood method) has been removed ; please use rather fastDNAml, which is much faster and equivalent.
  - Proteins
    protdist [advanced form] [protdist.doc]
    Distances from protein sequences.
    protpars [advanced form] [protpars.doc]
    Parsimony method for protein sequences.
- **Programs for distance matrix data [distance.doc]**

  neighbor [advanced form] [neighbor.doc]
  Neighbor-joining and UPGMA methods
  fitch [advanced form] [fitch.doc]
  Fitch-Margoliash and least-squares methods
  kitsch [advanced form] [kitsch.doc]
  Fitch-Margoliash and least-squares methods with molecular clock

Figure 13-7: Phylip online at the Pasteur Institute.

- Some methods enable you to measure pairwise distances on DNA or on proteins. There are also different programs for closely or distantly related sequences.
- Protdist is a method that uses a substitution matrix to measure the distance between your aligned sequences.
- Protpars does the same thing, but it tries to evaluate the number of mutations at the DNA level required for the observed amino-acid differences. Protpars works better with closely related sequences.

3. **Enter your e-mail address in the Your E-Mail field of the Phylip: Protdist page.**

   Again, a little briefing is in order:

   - If the server is overloaded or if you are analyzing a large alignment, Phylip sends you your results by e-mail rather than giving them to you online.
   - The e-mail you receive contains a URL. When you open this URL, your browser takes you to a page similar to what's shown in Figure 13-10.

- Do not give bogus e-mail addresses to the server! As soon as it times out, the server decides to return your results by e-mail. You cannot control this and if your e-mail was bogus, you will never know what your tree looked like! In our experience, it never took more than a few minutes before receiving the e-mail.

4. **Paste your multiple alignment into the Sequence window, as shown in Figure 13-8.**

   Do not forget to copy the header line that starts with CLUSTALW along with your alignment.

5. **Click the Bootstrap options link.**

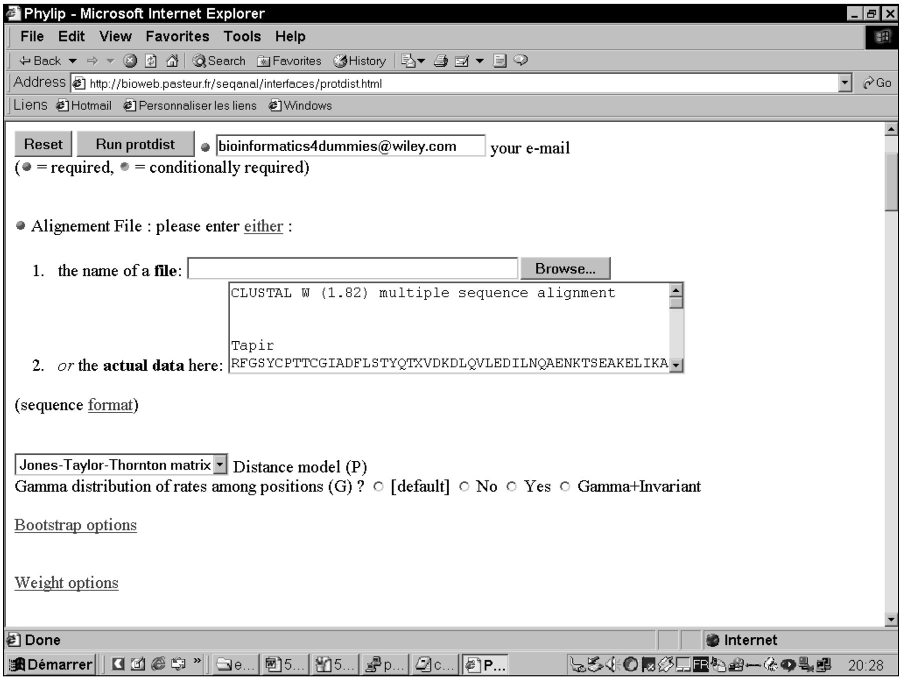
   You are brought to the Bootstrap Options section of the Web page.

6. **Select the Perform a Bootstrap Before Analysis check box, as shown in Figure 13-9.**

   Bootstrapping your data is a way of assessing the quality of your tree; see the "What is bootstrapping?" sidebar for details.

7. **Set the seed to some odd number such as *1* or *3*.**

   The *seed* controls the generation of random numbers during the bootstrap. Different seeds lead to different results.


Phylip - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://bioweb.pasteur.fr/seqanal/interfaces/protdist.html
Go
Liens Hotmail Personnaliser les liens Windows
Reset
Run protdist
Browse...
Done
Internet
Démarrer
20:28

● bioinformatics4dummies@wiley.com your e-mail
(● = required, ● = conditionally required)

● Alignement File : please enter either :

1. the name of a **file**:
2. *or* the **actual data** here:

```
CLUSTAL W (1.82) multiple sequence alignment


Tapir
RFGSYCPTTCGIADFLSTYQTXVDKDLQVLEDILNQAENKTSEAKELIKA
```

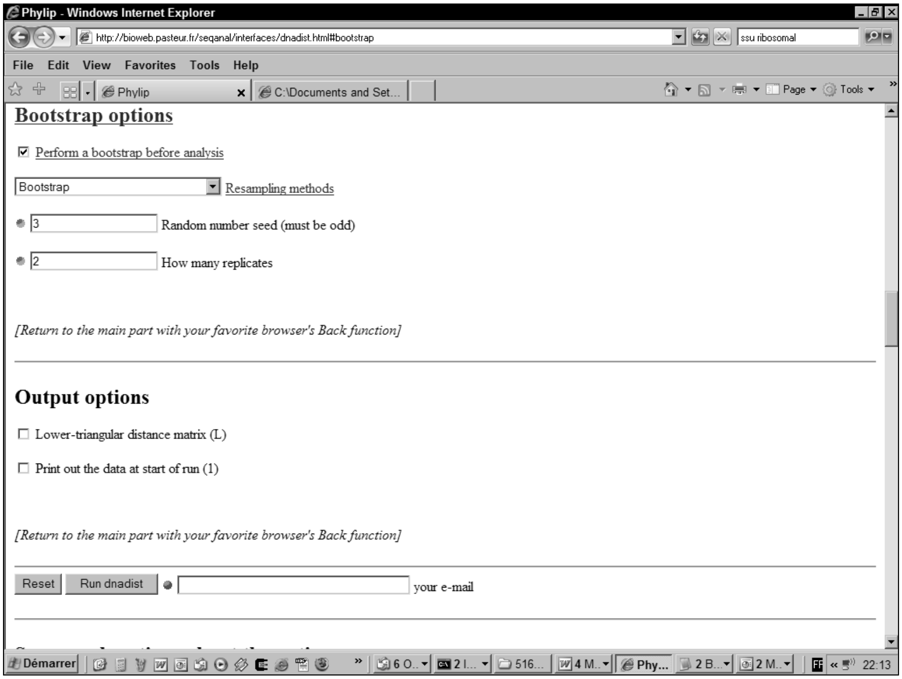
(sequence format)

Jones-Taylor-Thornton matrix Distance model (P)
Gamma distribution of rates among positions (G) ? ○ [default] ○ No ○ Yes ○ Gamma+Invariant

Bootstrap options

Weight options

Figure 13-8: Protdist server.


Phylip - Windows Internet Explorer
http://bioweb.pasteur.fr/seqanal/interfaces/dnadist.html#bootstrap
ssu ribosomal
File Edit View Favorites Tools Help
Phylip
C:\Documents and Set...
Page
Tools
Démarrer
22:13

Bootstrap options

☑ Perform a bootstrap before analysis

Bootstrap ▾ Resampling methods

● 3 Random number seed (must be odd)

● 2 How many replicates

[Return to the main part with your favorite browser's Back function]

Output options

□ Lower-triangular distance matrix (L)

□ Print out the data at start of run (1)

[Return to the main part with your favorite browser's Back function]

Reset | Run dnadist | ● your e-mail

Figure 13-9: The Bootstrap options of protdist.

8. **Keep the number of replicates to *100*.**

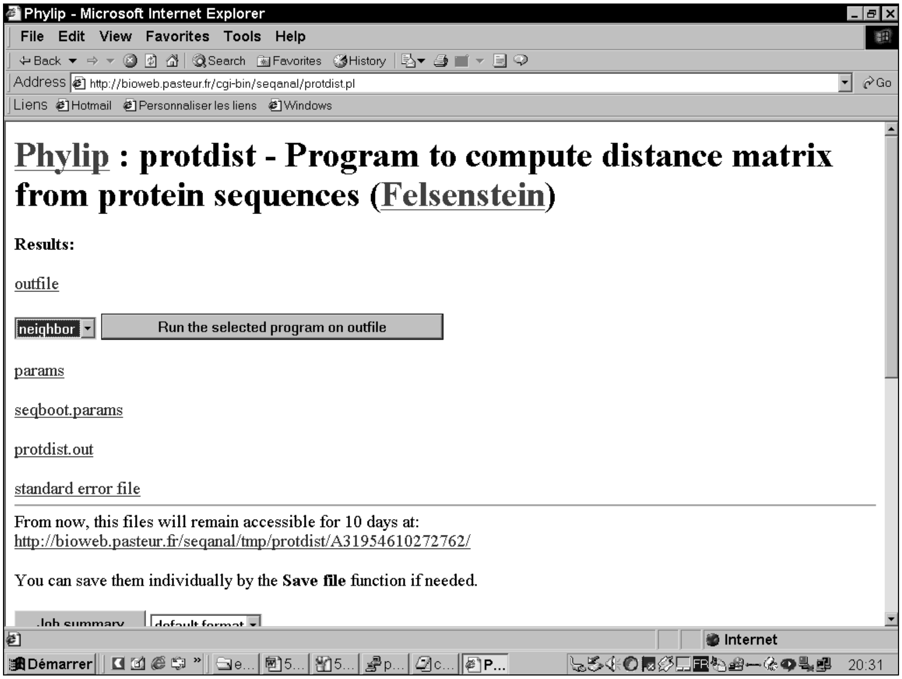
   The number of replicates is the number of bootstrap cycles that you do. Normally, this number should be at least 100.

   Remember how many replicates you request as you will need to re-enter this figure later!

   For real analysis, you need at least 100 replicates, but if you only want to get a feeling of how the server works, you may try values as small as 2. If you do so, it will be fine for the computation but don't expect the bootstrap to be biologically meaningful! For the purposes of this tutorial, we will use 2, so that the results get computed interactively, but if you require a large number of replicates, the server usually prefers returning the results by e-mail.

9. **Scroll back to the top of the page and click the Run Protdist button.**

   Eventually, you'll see results similar to what's shown in Figure 13-10. If your browser times out, the server will send you an e-mail that contains a URL. If you open this URL, you're taken to a page of results similar to what's shown in Figure 13-10.


Phylip - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History
Address http://bioweb.pasteur.fr/cgi-bin/seqanal/protdist.pl
Go
Liens Hotmail Personnaliser les liens Windows
neighbor
Run the selected program on outfile
Job summary
default format
Internet
Démarrer
20:31

Phylip : protdist - Program to compute distance matrix from protein sequences (Felsenstein)

Results:

outfile

params

seqboot.params

protdist.out

standard error file

From now, this files will remain accessible for 10 days at:
http://bioweb.pasteur.fr/seqanal/tmp/protdist/A31954610272762/

You can save them individually by the **Save file** function if needed.

Figure 13-10: Protdist results.

10. **Choose Neighbor from the first drop-down menu on the result page.**

    Neighbor is a program that turns your distance matrix into a Neighbor Joining tree.

    If you want, you can use another tree-reconstruction method (such as Fitch, which uses the mean-least-square method). These days, people usually agree that NJ provides a very good trade-off between available methods.

11. **Click the Run the Selected Program on Outfile button.**

    The Neighbor page appears.

12. **Select the type of tree you want to produce on the line that says Distance Method. (Neighbor Joining should be selected by default.)**

    Neighbor Joining and UPGMA are related methods, but Neighbor Joining produces trees that are much more accurate.

    UPGMA produces rooted trees, where the root is an attempt to guess the position of the common ancestor (See Figure 13-14).

    Neighbor-Joining does not try to guess where the common ancestor lies and produces unrooted trees that you can root later.

13. **Select the** Bootstrap options **link.**

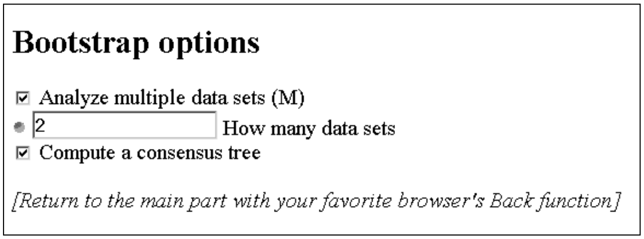
This takes you to a lower section of the page where you will enter some information related to the bootstrap we did in Figure 13-9.

This section is NOT identical to the Bootstrap section visited in Steps 5, 7, and 8. In Steps 5, 7, and 8 you declared how many Bootstrap cycles you wanted to do. Now that the work is done, you must remind the program how many cycles it has done!

14. **Select the Analyze Multiple Data Sets option.**

15. **Indicate the number of data sets you're providing, as shown in Figure 13-11.**

    Enter 100 if you have entered 100 in Step 8, or any number you have used in that step.



Bootstrap options

☑ Analyze multiple data sets (M)

● 2 How many data sets

☑ Compute a consensus tree

[Return to the main part with your favorite browser's Back function]

Figure 13-11: The Bootstrap option of Neighbor.

16. **Set the Random Seed to an odd number (such as 1 or 3), and check the Randomize box.**

    No need to use the same number as in Step 7.

17. **Check the Compute a Consensus Tree check box.**

    This option tells Neighbor to return the tree that has the best agreement with the replicate data sets you are providing.

    This option makes sense only if you are using the bootstrap and have chosen values larger than 1 in Steps 8 and 15 (multiple datasets).

18. **Scroll down to the Other Parameters section and choose an outgroup.**

    Here are some handy tips about outgroups:

    - Choosing an outgroup only matters if you selected Neighbor Joining in Step 10.
    - You can specify which outgroup you want to use according to its position in the multiple alignment, and Neighbor uses this outgroup as a root.
    - In your tree, the *root* represents the common ancestor of all the sequences. (Remember, however, that living fossils do not exist!) This outgroup is not the real root, but the real root is necessarily between the outgroup and its closest node.

Depending on the server and the default options, ClustalW sometimes changes the order of your sequences when it aligns them. It is, however, the order in the multiple alignment that prevails when you specify your outgroup. We therefore suggest that you take a look at your alignment to ensure that you're choosing the right outgroup.

19. **Scroll to the top of the page and click the Run Neighbor button.**

    Your results appear in a new page as a series of hyperlinks.

20. **Open the consensus tree files.**

    The two first files on the output list are the consensus tree files. They're hyperlinked, so you can simply click them to display their contents.

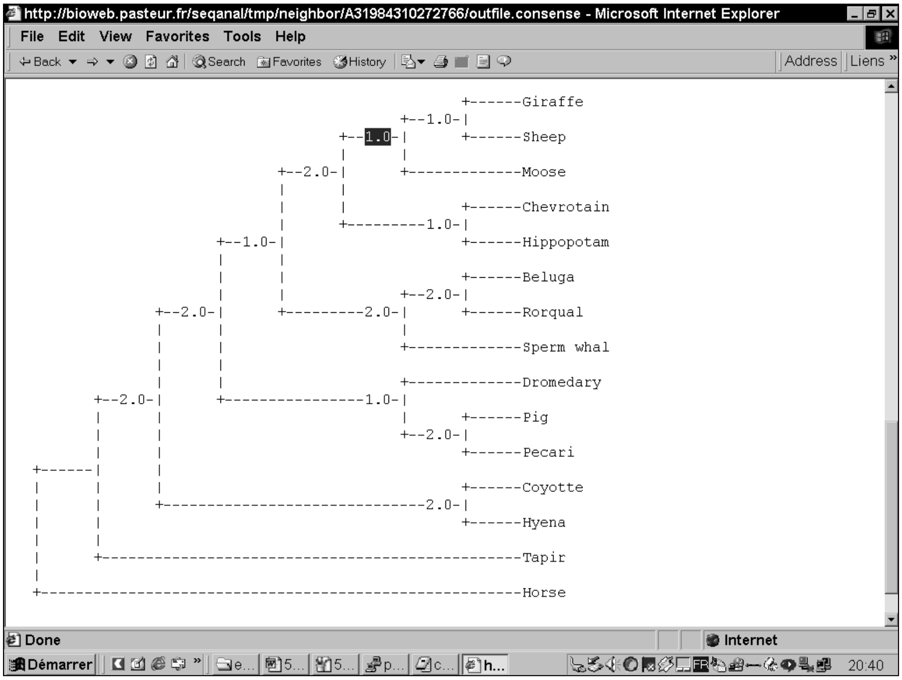
    `outfile.consense` (see Figure 13-12) contains a text version of your consensus phylogenetic tree.

    The consensus tree is always unrooted. The number above each branch indicates its stability, with 50 meaning more stable and 20 meaning less stable.

    `outtree.consense` is the same tree but in the Newick format.

21. **Open the normal trees.**

    Neighbor also outputs the trees it used to build the consensus in `outfile` and `outtree`.


http://bioweb.pasteur.fr/seqanal/tmp/neighbor/A31984310272766/outfile.consense - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back
Search
Favorites
History
Address
Liens
Done
Internet
Démarrer
20:40

```
                                                 +------Giraffe
                                          +--1.0-|
                                   +--1.0-|      +------Sheep
                                   |      |
                            +--2.0-|      +-------------Moose
                            |      |
                            |      |             +------Chevrotain
                            |      +---------1.0-|
                     +--1.0-|                    +------Hippopotam
                     |      |
                     |      |                    +------Beluga
                     |      |             +--2.0-|
              +--2.0-|      +---------2.0-|      +------Rorqual
              |      |                    |
              |      |                    +-------------Sperm whal
              |      |
              |      |                    +-------------Dromedary
       +--2.0-|      +----------------1.0-|
       |      |                           |      +------Pig
       |      |                           +--2.0-|
       |      |                                  +------Pecari
+------|      |
|      |      |                                  +------Coyotte
|      |      +------------------------------2.0-|
|      |                                         +------Hyena
|      |
|      +------------------------------------------------Tapir
|
+-------------------------------------------------------Horse
```

Figure 13-12: A consensus tree output by Neighbor.

- `outfile` contains a text representation of your trees. It contains as many trees as you requested replicates for in Steps 8 and 16.
- If you checked Neighbor Joining in Step 12, the trees are rooted with the outgroup.
- If you requested UPGMA, the trees have been automatically rooted.
- `outtree` contains the trees in Newick format.

22. **Save and re-display your output.**

    It is difficult to display a tree with bootstrap values using online tools. If you want to keep these values, the best thing to do is to save the `outfile.consense` file into a tree. (This is a standard text file that you can save by using your browser's File⇨Save As option.)

    If you want to redisplay your tree using another graphic format, the most convenient thing to do is to save the file `outtree.consense` and to turn it into a GIF or a PDF file with the Phylodendron server. (See the last part of this section for more on how to do this.)

What is bootstrapping?

Just because your computer can produce a tree doesn't necessarily mean that the tree is correct. Many things can go wrong in the tree's construction — including faulty data or an incorrect alignment. In most cases, your tree will be globally correct but contain a few inaccurate branches. This is where *bootstrapping* comes in handy. Tree bootstrapping is a way to check whether your tree is biologically meaningful by assessing its *robustness* — that is, checking whether every portion of the alignment equally supports your tree. The logic behind bootstrapping is that you measure the solidity of each branch in your tree so as to determine whether you can trust it when you draw your conclusions.

Bootstrapping is a multiple-step process. The program starts by sampling columns in your initial alignment. The purpose is to generate a *bootstrap alignment* in which some columns are missing and others are duplicated. Here's how it's done:

1. **Choose a column at random from the initial multiple alignment obtained with your sequences.**

   This column is the first column of your *bootstrap alignment.*

2. **Select a new column in the initial alignment and copy it to extend the bootstrap alignment.**

   By *extend* here, we mean making the alignment one column longer than the column you just copied from the initial alignment.

3. **Repeat Step 2 until your bootstrap alignment contains as many columns as your initial alignment.**

   Note that some of the columns are either never selected or appear several times in the final bootstrap alignment. This is *random sampling.*

(continued)

(continued)

For instance, here is how you could use an initial alignment to generate two bootstrap alignments:

```
Initial Alignment
Column 1 2 3 4 5 6 7 8 9
seq1   A B C D E F G H I
seq2   A A B B C B A C A
seq3   C C A C B A C A B

Bootstrap Alignment 1
1 1 8 1 2 5 1 8 2
A A H A B E A H B
A A C A B C A C A
C C A C C B C A C

Bootstrap Alignment 2
1 4 5 6 6 3 4 1 7
A D E F F C D A G
A B B C C B B A A
C C B A A A C C C
```

Note that the first bootstrap alignment contains Column #1 of the initial alignment four times.

This procedure generates many multiple alignments that look more or less like the original. The purpose of all this is to check whether all the columns in your initial alignment tell a similar evolutionary story (that should be the case).

During the next step, each random alignment is turned into a distance matrix, and each matrix is turned into a tree. To build the consensus tree, Phylip takes the average of all the trees it has generated from the bootstrap alignments. To assess the quality of each branch in the consensus tree, Phylip counts the number of bootstrap trees that contain this branch. Good branches are those that appear in every bootstrap tree. They are the ones on which you can strap your swing!

Take a look at the consensus tree back in Figure 13-12. On this consensus tree, every branch comes along with a number between 1 and 2. This value tells you how solid your branch is. This value falls between 1 and 2 because, in Steps 8 and 15, we chose to have 2 bootstrap cycles. Neighbor generated 2 bootstrap trees with these 2 bootstrap alignments, and it generated a consensus tree that's the average of these 2 bootstrap trees.

In a tree, a branch always separates your data into two groups: sequences on the left and sequences on the right side of the branch. The numbers above the branches in the consensus tree indicate how many branches exist in your two trees that split the data in exactly the same way as the branch you're looking at.

For instance, if you look at the consensus tree in Figure 13-12, you find that the branch containing giraffe and sheep has a value of only 1. The reason for this is that in one of the bootstrap trees, giraffe and moose are in the same group — while in the other tree, giraffe is in the same group as sheep (and moose is alone). In theory, you could conclude that the tree indicates some uncertainty on whether sheep and giraffe are more closely related to one another than each is related to moose. In practice, however, with only 2 bootstraps, you really can't say anything of the kind! You need at least 100 bootstrap cycles before you start convincing farmers to breed giraffes for their wool.

Making a maximum likelihood tree with PhyML

Maximum likelihood trees are considered to be more accurate than other trees because they produce the tree that is most likely (statistically speaking) to explain your alignment. In other words, your alignment is a little story that explains how, starting from one ancestral sequence, a series of mutations leads to all the observed diversity. Maximum likelihood is an attempt to ensure that the tree and the alignment tell exactly the same story.

It used to be that maximum likelihood trees were restricted to small datasets because they take too much time to compute. However, thanks to a recent breakthrough by Olivier Gascuel and Stéphane Guindon (Montpelier University, France) it is now possible to compute these trees in a much more realistic amount of time. We show you below how to use their PhyML server.

1. **Open the dummy alignment at** `www.tcoffee.org/dummy_aln3.html`**.**

   This alignment is in Phylip format, the only format recognized by the server.

2. **Use the File➪Save As option of your browser to save the alignment as a text file.**

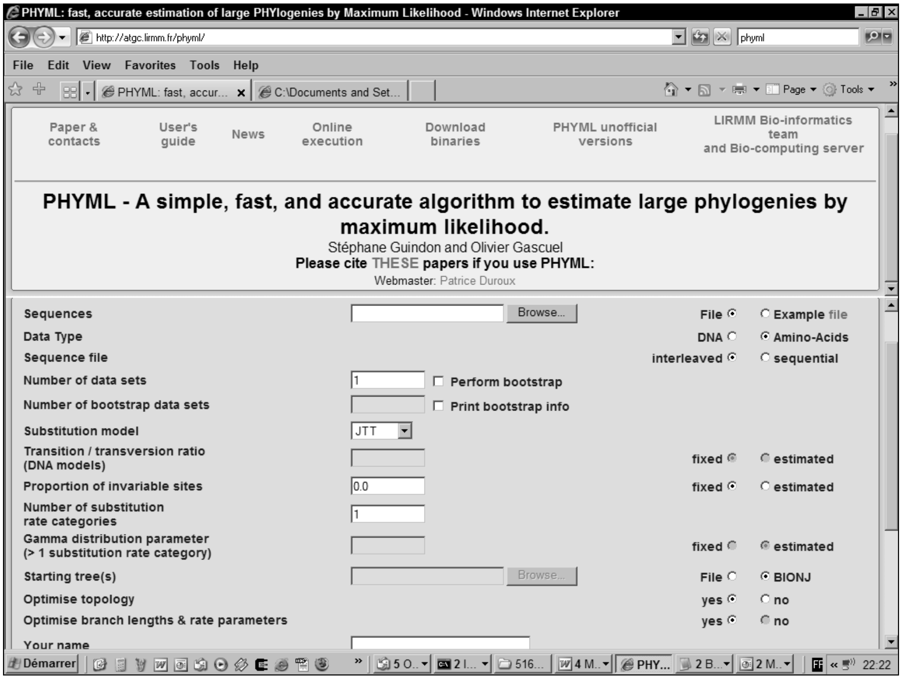
3. **Point your browser to** `atgc.lirmm.fr/phyml/`**.**

4. **Select the File radio button, as shown in Figure 13-13.**

   You can now upload your file with the Browse button.

5. **Select the Amino Acids radio button.**

   Our alignment contains amino acids.



PHYML - A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.

Stéphane Guindon and Olivier Gascuel

Please cite THESE papers if you use PHYML:

Webmaster: Patrice Duroux

Sequences	Browse...	File ●	○ Example file
Data Type		DNA ○	● Amino-Acids
Sequence file		interleaved ●	○ sequential
Number of data sets	1 □ Perform bootstrap		
Number of bootstrap data sets	□ Print bootstrap info		
Substitution model	JTT		
Transition / transversion ratio (DNA models)		fixed ●	○ estimated
Proportion of invariable sites	0.0	fixed ●	○ estimated
Number of substitution rate categories	1		
Gamma distribution parameter (> 1 substitution rate category)		fixed ○	● estimated
Starting tree(s)	Browse...	File ○	● BIONJ
Optimise topology		yes ●	○ no
Optimise branch lengths & rate parameters		yes ●	○ no
Your name			

Figure 13-13: Setting up the PHYML server.

6. **Select the Interleaved radio button.**

   Our alignment is in Phyllips Interleaved. This simply means the alignment is split into blocks (see the Chapter 9 for explanations of the various formats).

7. **Check the Perform Bootstrap checkbox and then enter 100 in the Number of Bootstrap Data Sets field.**

   Performing a bootstrap does take more time, but it is a crucial step if you want to reliably estimate the biological meaning of your tree.

   REMEMBER

   In practice, it's better to require 1,000 bootstraps — but for this example to run, we have set the value to 100.

8. **Scroll down to fill in the Name and E-Mail Address fields.**

9. **Click Run.**

When we did a trial run of this example, it took about half an hour for our results to come via e-mail — so be prepared for a bit of a wait. When an e-mail arrives, it contains a tree in Newick format that you can visualize using Phylodendron (we explain how to do this later in this chapter.) The tree will come with bootstrap values indicating how much trust you can have in each node. The higher the bootstrap value, the more reliable the node.

Knowing what's what in your tree

When you work with phylogenetic trees, there's always a bit of special vocabulary involved. Figure 13-14 recapitulates the names of the various components in your tree. Some of these names may sound a bit technical, but it's worth knowing their meaning because phylogenetic jargon is becoming more and more common in mainstream publications.

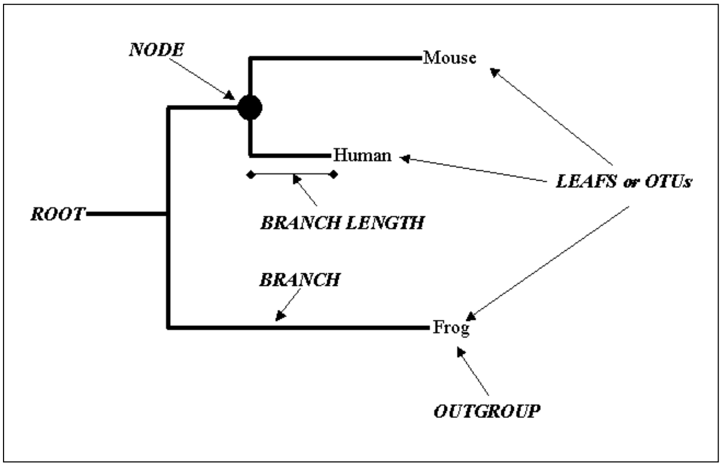


Figure 13-14: Various components of a phylogenetic tree.

Leaves are known as *OTUs* (Operational Taxonomic Units). Each *node* represents an ancestral OTU. A *clade* is a group of OTUs that includes several sequences and their common ancestor nodes. The *root* is the common ancestor of all the OTUs.

A *branch* defines the relation between a clade or an OTU and the rest of the tree. A tree can be *scaled* (in which case the branch length means something in terms of evolutionary time) or *unscaled* (in which case only the topology of the tree is informative).

The tree may be rooted or unrooted. *Unrooted* trees do not identify evolutionary paths. If your tree is unrooted, you accept the possibility that any of the nodes or OTUs could be closer to the common ancestor than to any of the other OTUs. If you have an *outgroup,* you know that the root must be inserted in the branch between this outgroup and the other clades.

REMEMBER

The root of a tree corresponds to the most ancient common ancestor. This sounds good, but it is (unfortunately) meaningless in terms of the kinds of trees we're reconstructing here. The problem is that tree-reconstruction methods have no way of telling the direction of evolution. They cannot be used to determine where the most ancient node is. As a consequence, the root you may observe on some trees is usually arbitrary — therefore meaningless from a biological point of view. Biologists like to use what they call *unrooted* trees, to avoid any confusion. Yet, if you need a root, the only proper way to insert it into your tree is to use what gurus call an *outgroup* — an organism so distantly related to the others that it can safely be used as a root. For instance, if you make a tree involving primates, you should include the mouse (with the exception of Mickey Mouse, few mice are considered primates) so you can use it as an outgroup; likewise, if you are comparing mammals, you should include a reptile, and so on.

Displaying your phylogenetic tree

Phylodendron is a powerful Web-based server that lets you turn a text file in Newick format into a graphic display of your tree. You can control every minute aspect of the display and generate images in any format you fancy. In the following steps, we show you how to use Phylodendron to turn a phylogenetic tree into a GIF picture that you can include in any document:

1. **Point your browser to** `iubio.bio.indiana.edu/treeapp/treeprint-form.html`.

   A similar service runs on the Spanish EMBnet node at `www.es.embnet.org/Doc/phylodendron/treeprint-form.html`.

2. **Paste your tree in Newick format in the Input window.**

WARNING!

3. **On the top row under Tree Styles, select the Phenogram radio button.**

   If you select Tree Diag, you must make sure you're providing the program with an unrooted tree; otherwise the program will hang.

4. **In the Extra Options section, choose GIF image from the Format drop-down menu.**
5. **If you want horizontal distances to have a meaning, select the Use Node Lengths check box in the Tree Growth section.**
6. **Click the Submit button.**
7. **Choose File➪Save As from your browser's main menu to save your results as a GIF file that you can use for inclusion in other documents.**

Doing Phylogeny for Free over the Internet

Conducting phylogeny over the Internet isn't yet as developed as database research or other kinds of sequence analysis methods. Fortunately, over the last few years, the community has made some high-quality resources available. These resources might not run on very powerful servers, but they're perfectly designed if all you need to do is build a tree once in a while.

Finding online resources

A vast amount of phylogeny resources are available over the Internet. Yet most of these programs aren't available online; you need to install them on your own computer. Table 13-2 lists a few resources that you can use online.

Table 13-2 Online Sites for Making Phylogenetic Trees

Address	*Description*
`www.ebi.ac.uk/clustalw/`	You can use ClustalW to build multiple alignments and compute NJ trees. Remember: You cannot do both at the same time!
`www.genebee.msu.ru/clustal/basic.html`	The Genebee server can produce genuine phylogenetic trees in one step.
`www.tcoffee.org`	Tcoffee computes a genuine NJ-phylogenetic tree in one step.

Address	Description
`www.jalview.org`	You can use Jalview to produce NJ trees. It's a very powerful tool that combines alignment editing with tree computation.
`atgc.lirmm.fr/phyml/`	A powerful method to compute maximum likelihood trees from Gascuel and his team.
`bioweb.pasteur.fr/seqanal/interfaces/bionj-simple.html`	An interface to BioNJ, a novel NJ method.
`www.up.univ-mrs.fr/evol/figenix/`	A powerful Java tool to gather members of a protein family and build the associated tree.
`bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`	A Web interface for Phylip.
`www.genebee.msu.ru/services/phtree_reduced.html`	Very powerful interface for a novel tree-reconstruction method.

Finding generic resources

Table 13-3 lists some of the most important resources in phylogeny. In most situations, you should be able to find what you need a few clicks away from these pages. We also include two resources that should cover your immediate need if you want to find out the basics of phylogeny for yourself.

Table 13-3 Generic Phylogenetic Resources on the Internet

Address	Description
`evolution.genetics.washington.edu/phylip/software.html`	Joe Felsenstein's pages, where Phylip lives; it's also one of the most extensive collections of resources available. Truly a legendary site!
`www.ucmp.berkeley.edu/subway/phylo/phylosoft.html`	A very complete list of phylogeny resources.
`paup.csit.fsu.edu/index.html`	The home of PAUP, legendary phylogeny package using Parsimony. Although PAUP is a commercial package, it's reasonably priced and worth every penny, according to specialists.

(continued)

Table 13-3 *(continued)*

Address	*Description*
`www.ncbi.nlm.nih.gov/About/primer/phylo.html`	The NCBI primer on phylogeny.
`www.techfak.uni-bielefeld.de/bcd/Curric/MathAn/mathan.html`	A high-quality course on tree reconstruction methods.

Collections of orthologous genes

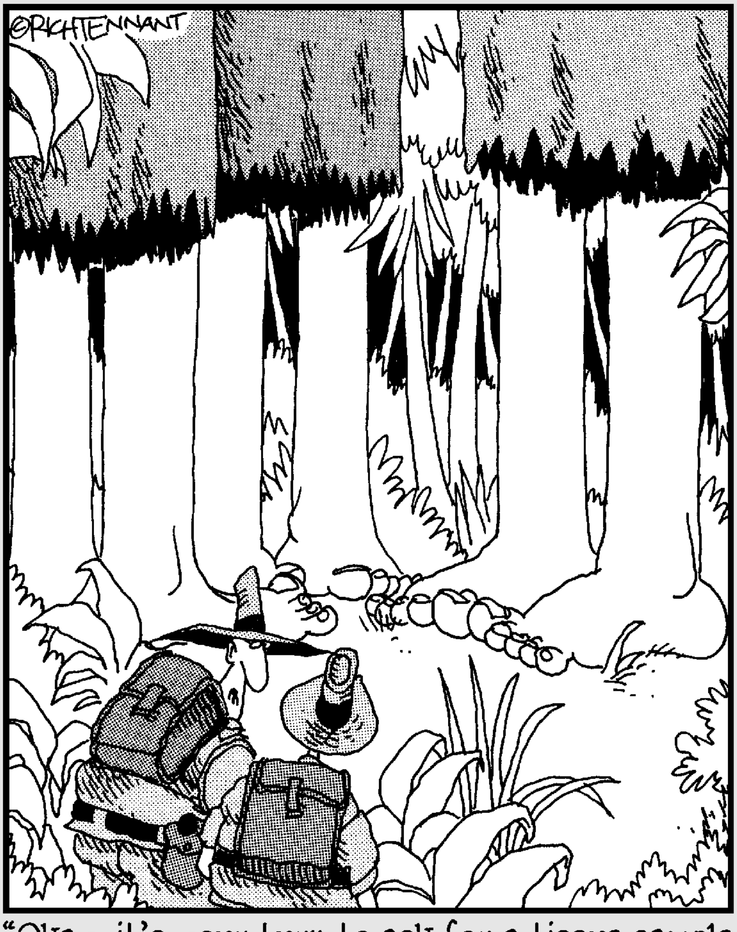
Collections of orthologous or homologous genes are very useful if you want to ask functional or phylogenetic questions. Table 13-4 lists some databases that provide this type of information.

Table 13-4 Collections of Orthologous Sequences

Address	*Description*
`www.ncbi.nlm.nih.gov/COG/`	Cluster of orthologous sequences maintained by the NCBI. Each cluster contains proteins from bacterial genomes.
`pbil.univ-lyon1.fr/databases/hovergen.html`	A collection of orthologous vertebrate genes.
`pbil.univ-lyon1.fr/databases/hogenom.html`	A collection of orthologous bacterial genes.
`systers.molgen.mpg.de`	Another collection of homologous sequences.
`rdp.cme.msu.edu/`, `www.psb.ugent.be/rRNA/ssu/`, `www.psb.ugent.be/rRNA/lsu/`	Three extensive collections of ribosomal RNA sequences, which are very useful for classifying new organisms and come with appropriate phylogenetic tools.

Part V
The Part of Tens

The 5th Wave By Rich Tennant



"Okay, it's your turn to ask for a tissue sample for DNA analysis."

In this part . . .

This part is just what it says on the box: It contains tables of links, resources, and various things useful but difficult to classify. If you haven't yet found what you were looking for anywhere else in this book, you'll probably find it here, or a couple of mouse clicks away from one of the addresses we give you. In this part, we also give you some useful guidelines if you are planning to let your research depend on Internet resources. Better read this before it's too late!

Chapter 14

The Ten (Okay, Twelve) Commandments for Using Servers

In This Chapter

- Making sure you don't divulge confidential data
- Ensuring that you're able to reproduce your work
- Using appropriate data
- Making sure you save the right files

I never did give anybody hell. I just told the truth, and they thought it was hell.

— Harry S. Truman (1884–1972)

The nicest thing about using Web-based servers is that (from the user's point of view) nothing in bioinformatics is ever a problem on the Internet. Consider: You can run any program you fancy. You don't need to update or maintain any database, and, if the server is broken, you can simply go and use another server. On the Web, you've got it so easy that you might even forget what you're doing and keep pushing buttons until something nice pops onto your screen. It's a bit like what you — and we — do in front of the television!

That's all fine and dandy, but if you need to use these bioinformatics tools for doing some research, you need to be a bit more careful, or this virtual paradise will rapidly turn into real hell. Nothing is more frustrating than a good result you can't reproduce, especially in science.

In this chapter, we give you a few pieces of advice designed to make life easy for you (in the online bioinformatic realm, anyway) if you want to produce results that your work can really depend on.

Keep in Mind: Your Data Is Never Secure on the Web

When you submit a sequence to an Internet server, you send it out into the world. With the exception of a few Java applets that you can use offline, never treat an Internet submission as a safe process.

While the people who run these servers are renowned for their honesty, they rarely have the means to ensure tight security for the data you send them. This means your data may be intercepted during transit.

We personally think that nobody really gives a damn about the sequences you send over the Internet. However, your employers probably think that the risk is real — and if you want to keep your job, you'd better share this fear!

If you need secrecy, you'll just have to install the programs you need on your own computer. No ifs, ands, or buts possible here.

Remember the Server, the Database, and the Program Version You Used

If you run the same program (such as ClustalW) on two different servers, it does not at all mean you're doing the same thing twice. The setup of the servers — or even the program version — can be different. This is the reason why *"I ran ClustalW over the Internet"* doesn't give other researchers enough information to reproduce your results — and it doesn't sound very professional, either.

To avoid any confusion, write down the name of the server as well as the version of the program that it runs. For instance, you should not expect a server that runs ClustalW 1.77 to return the same results as a server that runs ClustalW 1.81.

If your program (such as BLAST) uses a database (such as Swiss-Prot), you should write down the name and the version of the database you're searching.

Servers also change with time. They get upgraded, improved, and so on. On average, this happens every six months.

Write Down the Sequence-Identification Numbers

When you analyze a sequence, make sure that you have enough information on this sequence. Try to keep its ID (identification) and its AC (Accession) numbers, as well as any other information that seems to be relevant. Remember that the ID of an entry can change from one database release to the next, but the AC number never changes.

Write Down the Program Parameters

When you run a server, you often need to slightly alter the default parameters so the program does exactly what you need it to do. Many programs supplement their output with a file that contains all the parameters you used. This file often has a name that contains the word *log.* For BLAST, these parameters are at the bottom of the result page. If your program outputs such a file, be sure to save it.

As an alternative, you can use the PrntScrn facility of Windows to make a screen shot. (See the next commandment, “Save Your Internet Results the Right Way,” for tips on how to do this.) The PrntScrn technique can ease the process of making an exact reproduction of your request.

Save Your Internet Results the Right Way

Saving Internet results can be trickier than it seems at first glance. This is especially true when you try to save documents that contain images like BLAST results. We cannot give you a foolproof procedure because what works best changes all the time — depending on your Web browser, the way the browser is installed on your computer, and the kind of data you’re downloading. When it comes to saving files, Java is a major source of trouble, and often the only solution is to take a screen shot.

WARNING!

Never assume that an important result has been saved correctly; always double-check. The latest version of Internet Explorer and Netscape makes it possible to save complete pages (including pictures).

If you can't save your page, try to print it as a PDF file.

If printing into a PDF file fails as well, make a screen shot by using the Windows PrntScrn utility:

1. **Press the PrntScrn key on your keyboard.**

   This puts a copy of your screen onto the Clipboard.
2. **Open Microsoft Paint by choosing Start➪Programs➪Accessories➪Paint.**
3. **Press Ctrl+V and say Yes to the message that proposes that you resize the bitmap.**
4. **Save as usual or print.**

   Save using the `.jpg` format to generate a compressed picture.

Whenever possible, also save a text version of your results! Many programs output a staid text version of the results along with the flashy graphic output. This text version is a much easier way to store and exploit your results. In many cases, it is the only format that other programs can use.

Use E-Values

E-values (expectation values) tell you how many times a result as good as the one you're looking at could have been reached by chance alone. On its own, this has no special biological meaning. However, the rationale behind the use of E-values in biology is that a good result is something that chance could not have produced without lots of help from nature.

A good E-value is a low E-value. For instance, an E-value of 10-^{32} is better than an E-value of 10-^{4}.

Make Sure You Can Trust Your Alignments

Just because you have an alignment doesn't mean you can trust it implicitly! Before you build an elaborate theory on your alignment, make sure it's biologically correct. Measure the percentage of identity and look for conserved blocks or important conserved residues. Also, be critical and use the guidelines we give you in Chapter 9!

Use Different Programs to Check Borderline Results

If you're not sure of your results, try to reproduce them by using different methods. If you can, it probably means that your results are correct and meaningful. If you cannot, it could mean that your results are only a collection of meaningless artifacts. For instance, if you generate a phylogenetic tree with two methods and you get two very different trees, there's a very good chance that none of these trees is, in fact, the correct one. Being able to reproduce the same results with different methods is something biologists sometimes name *robustness,* and it is a notion that applies to most of what we said in this book. (See the identification of transmembrane domains in Chapter 6, for instance.)

When you use different programs, make sure that you double-check with a method that relies on a different principle. For instance, there's no point in comparing an NJ tree from ClustalW with another NJ tree from Phylip. What you want to do is confront your NJ tree with a Fitsch or ML tree.

As an alternative, you can also change some of the key parameters of your program and check how they affect the results you're interested in. In a multiple alignment, for instance, you could change the gap penalties.

Stay Away from Unpublished Methods!

This isn't to say that you should never try something new, but using well-established methods is the best way to produce results that people trust. Never use an unknown program that's not published and whose principle you don't understand. In brief, stay away from any student's midterm-project pages!

Databases Are Not Like Good Wine

Databases are not like good wine: They do not age well! Avoid using data that a colleague downloaded last year and left to rot away on a PC.

Any time you need a database or some sequence data, always make sure that you're using fresh data.

Just Because It Looks Free Doesn't Mean It Is Free . . .

Public data isn't always available for you to use anyway you see fit. Before you start using some data on a server, make sure that you know whom this data belongs to. If you work in a company, you may have to purchase a license before using it.

WARNING!

Many companies advertise themselves by providing free data or services. This lasts only as long as the company's management types see fit. If your work depends on such resources, an abrupt discontinuation can mean disaster! We aren't saying that you should never use private resources, but you should always know how long those resources are going to last.

Biting the Bullet at the Right Time

The reason we've dedicated this entire book to online resources is because online resources are easy to use and don't require any kind of complicated setup. They let you concentrate on biology, not on your computer! This is exactly what you need . . . if you use bioinformatics only occasionally. Don't expect to be able to assemble and annotate your own viral or bacterial genome, however, without some serious investment in bioinformatics — and access to knowledgeable people!

If you find that you start relying heavily on some servers, you'll have to bite the bullet and install these programs on your own computer. Installing and maintaining software is easier than it seems — such software, however, works best on Linux or Unix systems. If you start from scratch, learning all you need to know to install and use Linux would probably take you a couple of weeks. Even if your time is tight, you can consider that a safe investment.

How will you know when that time has come? Well, if you find yourself doing 20 BLASTs in a row using an online server, it's a sure sign that you need the independence and flexibility that comes with having the software installed on your own computer!

Chapter 15

Some Useful Bioinformatics Resources

In This Chapter

- Ten important databases
- Some very important programs
- Our favorite link collections and generic resources
- Finding out about the latest techniques

> *"How did you know that I knew you knew?"*
>
> — Typical example of knowledge-based programming

One of the very nice and unique things about bioinformatics is that (almost) everything you need is available online. In this chapter, we show you around the main resources. If you have a fairly good idea of what's behind every link we list here, consider yourself fairly well-trained already in the field of bioinformatics.

Remember that this field is moving rapidly. Being operational in bioinformatics is not so much about knowing things but about knowing where to find the knowledge you need. This book is not exhaustive at all, but it offers you many gateways. Do not hesitate to go further and explore. While on this fascinating journey, keep new things in the corner of your mind until you need them.

Ten Major Databases

Bioinformatics is about exploring biological information. This information is safely kept in databases. These databases are our modern museums, cyber-reference collections where all the knowledge of an era is carefully stored and classified. Table 15-1 lists ten of these databases. It isn't at all an exhaustive list, but a biased vision that this chapter title limits to just ten items!

Table 15-1 Ten Important Bioinformatics Databases

Name	*Address*	*Has Data About*
GenBank/DDBJ/EMBL	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Nucleotide`	Nucleotide sequences
Ensembl	`www.ensembl.org`	Human/mouse genome
PubMed	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?DB=pubmed`	Literature references
NR	`www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=Protein`	Non-redundant protein sequences
Swiss-Prot	`www.expasy.ch`	Protein sequences
InterPro	`www.ebi.ac.uk`	Protein domains
OMIM	`www.ncbi.nlm.nih.gov`	Genetic diseases
Enzymes	`www.chem.qmul.ac.uk`	Enzymes
PDB	`www.rcsb.org/pdb/`	Protein structures
KEGG	`www.genome.ad.jp`	Metabolic pathways

Ten Major Bioinformatics Software Programs

Data is nothing if you don't know how to analyze it. This is what bioinformatics software programs are made for. You can use these programs to search for your data, to analyze it, and to display it. Table 15-2 lists the five main categories of programs that you may come across in bioinformatics, along with what we believe to be the most characteristic pieces of software in each of these categories. Each of these programs is introduced somewhere else in this book, as indicated in the table.

Table 15-2 also contains the address of an online server for each program. In most cases, you can find alternative addresses in the corresponding chapter.

Table 15-2 Ten Important Software Programs in Bioinformatics

Category	*Name*	*Address*	*Description*
Database Search	SRS	`srs.ebi.ac.uk`	Database search
	Entrez	`www.ncbi.nih.gov/Entrez`	Database search (Chapter 3)
	BLAST	`www.ncbi.nlm.nih.gov/blast`	Homology search (Chapter 7)
	DALI	`www.ebi.ac.uk/dali`	Structure database search (Chapter 11)
Multiple alignment	ClustalW	`www.ebi.ac.uk/clustalw`	Multiple sequence alignment (Chapter 9)
	MUSCLE	`phylogenomics.berkeley.edu/muscle/`	Multiple sequence alignment (Chapter 9)
	Tcoffee	`www.tcoffee.org`	Multiple sequence alignment (Chapter 9)
Prediction	GenScan	`genes.mit.edu`	Gene prediction (Chapter 5)
	PsiPred	`bioinf.cs.ucl.ac.uk/psipred/`	Protein structure prediction (Chapter 11)
	Mfold	`www.bioinfo.rpi.edu/applications/mfold/`	RNA structure prediction (Chapter 12)
Phylogenetics	Phylip	`bioweb.pasteur.fr/seqanal/phylogeny/phylip-uk.html`	Tree reconstruction (Chapter 13)
	PhyML	`atgc.lirmm.fr/phyml/`	Tree reconstruction (Chapter 13)
Edition/ Visualization	Jalview	`www.jalview.org`	Alignment editor (Chapter 10)
	Logos	`weblogo.berkeley.edu`	A MSA Visualization Tool (Chapter 10)
	Trees	`iubio.bio.indiana.edu/treeapp/treeprint-form.html`	Tree Visualization (Chapter 13)
	Rasmol	`www.umass.edu/microbio/rasmol/`	Structure visualization (Chapter 11)

Ten Major Resource Locators

Bioinformatics changes fast. State-of-the-art programs become obsolete, and established programs disappear, institutes are renamed, and new programs come out. That's the way of the world. If you find this ongoing change a bit unsettling, here's some reassuring news: Anything new you might be looking for is probably only one or two clicks away from the generic resources we list in Table 15-3. In this list, the last four links take you to some very good online resources where you can find almost anything you want on bioinformatics-related topics.

Table 15-3 Ten Bioinformatics Resource Locators

Name	*Address*	*Description*
ExPASy	`www.expasy.ch`	Dedicated to proteins
ArrayExpress	`www.ebi.ac.uk/microarray/`	DNA chips
Swbic	`www.swbic.org/`	Miscellaneous links
Pasteur	`bioweb.pasteur.fr/intro-uk.html`	Miscellaneous links; many online tools
RNA World	`www.imb-jena.de/RNA.html`	RNA-related links
miRNAs	`microrna.sanger.ac.uk/sequences/index.shtml`	Extensive Resources on miRNA
Phylip	`evolution.genetics.washington.edu/phylip/software.html`	Everything on phylogeny
NCBI primers	`www.ncbi.nlm.nih.gov/education`	Very good primers on many subjects
Bielefeld	`bibiserv.techfak.uni-bielefeld.de/intro/dist.html`	Awesome online course
Bio-informer	`www.ebi.ac.uk/Information/News/`	The EBI online news
Coffee Corner	`www.ncbi.nlm.nih.gov/books/bv.fcgi?call=bv.View..ShowSection&rid=coffeebrk`	NCBI Online News

Some Places to Find Out What's Really Going On

If you are always looking for fresh meat and just-out-of-the-oven cakes, the resources in Table 15-4 are for you. Better take a look if you want to turn into a guru yourself.

Table 15-4 Ten Places to Go Farther

Name	*Address*	*Description*
Nucleic Acid Research	`nar.oxfordjournals.org/`	Once a year, NAR publishes both a database issue and Web-server issue. These are available for free — and contain the state of the art in bioinformatics.
Bioinformatics	`bioinformatics.oxfordjournals.org/`	Bioinformatics contains articles describing the most recent methods in bioinformatics.
Conferences	`www.iscb.org/events/event_board.php`	An exhaustive list of major conferences in the field of bioinformatics, provided by the International Society For Computational Biology.

Index