# Bioinformatics and Biomarker Discovery

## "omic" data analysis for personalized medicine

Francisco Azuaje

WILEY-BLACKWELL

# Bioinformatics and Biomarker Discovery

# Bioinformatics and Biomarker Discovery

"Omic" Data Analysis for Personalized Medicine

**Francisco Azuaje**
*Public Research Centre for Health (CRP-Santé), Luxembourg*

WILEY-BLACKWELL
A John Wiley & Sons, Ltd., Publication

Wiley-Blackwell is an imprint of John Wiley & Sons, formed by the merger of Wiley's global Scientific, Technical and Medical business with Blackwell Publishing.

*Registered office*: John Wiley & Sons Ltd, The Atrium, Southern Gate, Chichester, West Sussex, PO19 8SQ, UK

*Other Editorial Offices:*
9600 Garsington Road, Oxford, OX4 2DQ, UK
111 River Street, Hoboken, NJ 07030-5774, USA

For details of our global editorial offices, for customer services and for information about how to apply for permission to reuse the copyright material in this book please see our website at www.wiley.com/wiley-blackwell



*Library of Congress Cataloguing-in-Publication Data*

Azuaje, Francisco.
Bioinformatics and biomarker discovery : "omic" data analysis for personalized medicine / Francisco Azuaje.
p. ; cm.
Includes bibliographical references and index.
ISBN 978-0-470-74460-4
1. Biochemical markers. 2. Bioinformatics. I. Title.
[DNLM: 1. Computational Biology. 2. Biological Markers. 3. Genomics–methods. 4. Statistics as Topic. QU 26.5 A997b 2010]
R853.B54A98 2010
610.285–dc22

2009027776

ISBN: 978-0-470-74460-4

A catalogue record for this book is available from the British Library.

Set in 10/12 Times by Thomson Digital, Noida, India
Printed in Singapore by Markono

1 2010

*To my family:*
*Alayne*
*Nelly and Francisco José*
*Nelytza, Oriana and Valentina*

# Contents

# Author and guest contributor biographies

Francisco Azuaje has more than fifteen years of research experience in the areas of computer science, medical informatics and bioinformatics. His contributions have been reflected in several national and international research projects and an extensive publication record in journals, conference proceedings and books. Dr Azuaje is a Senior Member of the IEEE. He held a lectureship and readership in computer science and biomedical informatics at Trinity College Dublin, Ireland, and at the University of Ulster, UK, from January 2000 to February 2008. He is currently leading research in translational bioinformatics and systems biology approaches to prognostic biomarker development at the Laboratory of Cardiovascular Research, CRP-Santé, Luxembourg. He has been a member of the editorial boards of several journals and scientific committees of international conferences disseminating research at the intersection of the physical and computer sciences, engineering and biomedical sciences. He is an Associate Editor of the IEEE Transactions on Nanobioscience and BioData Mining. Dr Azuaje co-edited the books: *Data Analysis and Visualization in Genomics and Proteomics*, *Artificial Intelligence Methods and Tools for Systems Biology*, and *Advanced Methods and Tools for ECG Data Analysis*. He is currently a Section Editor of the *Encyclopaedia of Systems Biology*.

**Guest contributor biographies**

Guest commentary on chapter 4

Ana Dopazo holds a PhD in Molecular Biology and has worked in the field of gene expression analysis for more than 16 years, including periods in the USA, Germany and Spain, both in academia and in private companies. She currently heads the Genomics Unit at the CNIC (Centro Nacional de Investigaciones Cardiovasculares) in Madrid. The CNIC Genomics Unit is dedicated to providing high-quality genomic technology as a key element in the expansion of our knowledge of genomes, mainly in the context of translational cardiovascular research. The Unit has extensive experience in the study of transcriptomes by means of DNA microarrays, and the group's current array-based studies include genome-wide gene (mRNA) and microRNA expression analysis and whole-genome microarray differential gene expression analysis at the exon-level. The Unit's expertise in array-based transcriptome analysis encompasses all steps required by these approaches, including experimental design, sample preparation and processing, and statistical data analysis.

Guest commentary on chapter 5
Haiying Wang received a PhD degree on artificial intelligence in biomedicine from the University of Ulster, Jordanstown, UK, in 2004. He is currently a lecturer in the School of Computing and Mathematics at the University of Ulster. His research interests include knowledge engineering, data mining, artificial intelligence, XML, and their applications in medical informatics and bioinformatics. Since 2000, he has published more than 50 publications in scientific journals, books and conference proceedings related to the areas at the intersection of computer science and life science.

Huiru Zheng (IEEE member) is a lecturer in the Faculty of Engineering at the University of Ulster, UK. Dr Zheng received a BEng degree in Biomedical Engineering from Zhejiang University, China in 1989, an MSc degree in Information Processing from Fuzhou University, China in 1992, and a PhD degree on data mining and Bioinformatics from the University of Ulster in 2003. Before she joined the University of Ulster, she was working in Fuzhou University, China, as an Assistant Lecturer (1992), Lecturer (1995) and Associate Professor (2000). Her research interests include biomedical engineering, medical informatics, bioinformatics, data mining and artificial intelligence. She has over 80 publications in journals and conferences in these areas.

Guest commentary on chapter 6
Kenneth Bryan graduated from Trinity College Dublin with a degree in Microbiology in 2001. He attained a Graduate Diploma in IT 2002 at Dublin City University before returning to Trinity College to complete a PhD in Machine Learning/Bioinformatics in 2006 which chiefly focused on bicluster analysis of microarray gene expression data. During 2006–2008 Dr Bryan worked as a post-doctoral researcher in the Machine Learning group in the Complex and Adaptive Systems Laboratory (CASL) in University College Dublin in a number of areas including semi-supervised classification of gene expression data, feature selection in metabolomics data and adapting bioinformatics metrics to alternative domains. In 2008 Dr Bryan joined the Cancer Genetics group at the Royal College of Surgeons, Ireland and is currently carrying out research into molecular events that lead to the development and progression of paediatric cancers, particularly Neuroblastoma.

Guest commentary on chapter 7
Zhongming Zhao received his PhD degree in human and molecular genetics from the University of Texas Health Science Centre at Houston, USA in 2000. He also received three MSc degrees in genetics (1996), biomathematics (1998), and computer science (2002). After completion of his Keck Foundation postdoctoral fellowship, he became an assistant professor of bioinformatics in the Virginia Commonwealth University, USA, in August 2003. He became an associate professor in the Department of Biomedical Informatics, Vanderbilt University, and Chief Bioinformatics Officer in

Vanderbilt-Ingram Cancer Center, USA in 2009. His research interests are bioinformatics and systems biology approaches to studying complex diseases (data management, integration, gene ranking, gene features and networks, etc.); genome-wide or large-scale analysis of genetic variation and methylation patterns; microRNA gene networks; comparative genomics; and biomedical informatics. He has published more than 50 papers in these areas. He served as editorial board member in six journals and program committee member and session chair in nine international conferences including WICB'06, BMEI'08, ICIC'08, IJCBS'09, and SSB'09. He received several awards, including the Keck Foundation Post-doctoral Fellowship (twice: 2002, 2003), White Magnolia Award (2006), NARSAD Young Investigator Award (twice, 2005, 2008) and the best paper award from the ICIC'08 conference.

Guest commentary on chapter 8
Yves Moreau is a Professor of Engineering at the University of Leuven, Belgium. He holds an MSc in Engineering from the Faculté Polytechnique de Mons, Belgium and an MSc in Applied Mathematics from Brown University, RI, where he was a Fulbright scholar. He holds a PhD in Engineering from the University of Leuven. He is co-founder of two spin-offs of the University of Leuven: Data4s (www.norkom.com) and Cartagenia (www.cartagenia.com), the last one being active in clinical genetics. His research focuses on the application of computational methods in systems biology towards the understanding and modulation of developmental and pathological processes in constitutional disorders. Thanks to a unique collaboration with the Centre for Human Genetics, University Hospitals Leuven, his team develops an integrative computational framework for supporting genetics research from patient to phenotype to therapy. From a methodological point of view, his team develops methods based on statistics, probabilistic graphical models, and kernel methods for such analyses, with an emphasis on heterogeneous data integration and the development of computational platforms that are directly useful to biologists.

Guest commentary on chapter 10
Gary B. Fogel is Chief Executive Officer of Natural Selection, Inc. (NSI) in San Diego, California. He joined NSI in 1998 after completing a PhD in biology from the University of California, Los Angeles, with a focus on the evolution and variability of histone proteins. While at UCLA, Dr Fogel was a Fellow of the Centre for the Study of Evolution and the Origin of Life and earned several teaching and research awards. Dr Fogel's current research interests focus on the application of computational intelligence methods to problems in biomedicine and biochemistry, such as gene expression analysis, gene recognition, drug activity/toxicity prediction, structure analysis and similarity, sequence alignment, and pattern recognition. Dr Fogel is a senior member of the IEEE and member of Sigma Xi. He currently serves as Editor-in-Chief for *BioSystems*, and as an associate editor for *IEEE Transactions on Evolutionary Computation* and *IEEE Computational Intelligence Magazine*. He co-edited a volume on *Evolutionary Computation in Bioinformatics*, published in 2003 (Morgan Kaufmann) and co-edited *Computational Intelligence in Bioinformatics*, published in 2008 (IEEE Press). Dr Fogel serves as conference chair for the 2010 IEEE Congress on Evolutionary Computation (http://www.wcci2010.org) held as part of the IEEE World Congress on Computational Intelligence.

Guest commentary on chapter 10

Riccardo Bellazzi is Associate Professor of Medical Informatics at the Dipartimento di Informatica e Sistemistica, University of Pavia, Italy.

He teaches Medical Informatics and Machine Learning at the Faculty of Biomedical Engineering and Bioinformatics at the Faculty of Biotechnology of the University of Pavia. He is a member of the board of the PhD in Bioengineering and Bioinformatics of the University of Pavia.

Dr Bellazzi is Past-Chairman of the IMIA working group of Intelligent Data Analysis and Data Mining, program chair of Medinfo 2010, the world conference on Medical Informatics and of the AIME 2007 conference; he is also part of the program committee of several international conferences in medical informatics and artificial intelligence. He is a member of the editorial board of *Methods of Information in Medicine* and of the *Journal of Diabetes Science and Technology*. He is affiliated with the American Medical Informatics Association and with the Italian Bioinformatics Society. His research interests are related to biomedical informatics, comprising data mining, IT-based management of chronic patients, mathematical modelling of biological systems and bioinformatics. Riccardo Bellazzi is author of more than 200 publications on peer-reviewed journals and international conferences.

# Acknowledgements

I thank my wife, Alayne Smith, for continuously helping me to succeed in personal and professional challenges. Her patience and understanding were essential to allow me to overcome the many obstacles encountered during the development of this project. The love and teachings given to me by my parents, Nelly and Francisco José, have been the greatest sources of support and inspiration for accomplishing my most valued contributions and aspirations. I thank my sister, Nelytza, and my nieces, Oriana and Valentina, for teaching me great lessons of personal strength, determination and compassion in the face of adversity.

Highly esteemed colleagues: Ana Dopazo, Haiying Wang, Huiru Zheng, Kenneth Bryan, Zhongming Zhao, Yves Moreau, Riccardo Bellazzi and Gary Fogel, enriched this project through the contribution of commentaries to accompany some of the chapters. I also thank them for their advice and corrections that allowed me to improve the content and presentation of this book.

I appreciate the support I have received from Fiona Woods and Izzy Canning, Project Editor and Publishing Assistant respectively, at John Wiley & Sons. I thank production staff at John Wiley & Sons for their assistance with book formatting and cover design. I also appreciate the help and advice from Andrea Baier during the early stages of this project. I thank Poirei Sanasam, at Thomson Digital, for management support during final production stage.

I thank all those colleagues and students, who over the years have helped me to expand my understanding of science and education. In particular, I express my affection for my school teachers, university mentors and friends in my homeland, Venezuela. Their experiences and generosity have greatly influenced my love for scientific knowledge and research.

# Preface

Biomarkers are indicators of disease occurrence and progression. Biomarkers can be used to predict clinical responses to treatments, and in some cases they may represent potential drug targets. Biomarkers can be derived from solid tissues and bio-fluids. Also they can refer to non-molecular risk or clinical factors, such as life-style information and physiological signals. Different types of biomarkers have been used in clinical practice to detect disease and predict clinical outcomes.

Advanced laboratory instruments and computing systems developed to decipher the structure and function of genes, proteins and other substances in the human body offer a great variety of imperfect yet potentially useful data. Such data can be used to describe systems and processes with diverse degrees of accuracy and uncertainty. These limitations and the complexity of biomedical problems represent natural obstacles to the idea of bringing new knowledge from the laboratory to the bedside.

The greatest challenge in biomarker discovery is not the discovery of powerful predictors of disease. Nor is it the design of sophisticated algorithms and tools. The greatest test is to demonstrate its potential relevance in a clinical setting. This requires strong evidence of improvements in the health or quality of life of patients. This also means that potential biomarkers should stand the challenge of independent validations and reproducibility of results.

Advances in this area have traditionally been driven at the intersection of the medical and biological sciences. Nevertheless, it is evident that current and future progress will also depend on the combination of skills and resources originating from the physical and computational sciences and engineering. In particular, bioinformatics and computational biology have the mission to bring new capacities and possibilities to understand and solve problems.

The promise of new advances based on the synergy of these disciplines will also depend on the growth and maturation of a new generation of researchers, managers and policy makers. This will be accomplished only through new and diverse training opportunities, ranging from pre-college, through undergraduate and post-graduate, to post-doctoral and life-long education.

One of the crucial challenges for bioinformaticians and computational biologists is the need to continuously accumulate a great diversity of knowledge and skills. Moreover, despite the fact that almost everyone in the clinical and biological sciences would agree on the importance of computational research in translational biomedical research, there are still major socio-cultural obstacles that must be overcome. Such obstacles mirror the complexity and speed of unprecedented changes in technology, scientific culture and human relations.

Bioinformaticians and computational biologists have a mission that goes beyond the provision of technical support or the implementation of standard computing solutions. Their mission is to contribute to the generation and verification of new knowledge, which can be used to detect, prevent or cure disease. In the longer term, this may result in a more effective fight against human suffering and poverty. This demands from us a continuous improvement of skills and changes in attitude. Skills and attitudes that can prepare us to cooperate and lead in this endeavour.

This book aims to support efforts in that direction. It represents an attempt to introduce readers to some of the crucial problems, tools and opportunities in bioinformatics and biomarker research. I hope that its content will at least serve to foster new conversations between and within research teams across disciplines, or even to help to recognize new value and purpose of ongoing interactions.

# 1 Biomarkers and bioinformatics

This chapter discusses key concepts, problems and research directions. It provides an introduction to translational biomedical research, personalized medicine, and biomarkers: types and main applications. It will introduce fundamental data types, computational and statistical requirements in biomarker studies, an overview of recent advances, and a comparison between 'traditional' and 'novel' molecular biomarkers. Significant roles of bioinformatics in biomarker research will be illustrated, as well as examples of domain-specific models and applications. It will end with a summary of expected learning outcomes, content overview, and a description of basic mathematical notation to be used in the book.

## 1.1 Bioinformatics, translational research and personalized medicine

In this book, the term bioinformatics refers to the design, implementation and application of computational technologies, methods and tools for making 'omic' data meaningful. This involves the development of information and software resources to support a more open and integrated access to data and information. Bioinformatics is also used in the context of emerging computational technologies for modelling complex systems and informational patterns for predictive purposes. This book is about the discovery of knowledge from human molecular and clinical data through bioinformatics. Knowledge that represents 'biomarkers' of disease and clinically-relevant phenotypes.

Another key issue that this book addresses is the 'translational' role of bioinformatics in the post-genome era. Translational research aims to aid in the transformation of biological knowledge into solutions that can be applied in a clinical setting. In addition,

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

this involves the incorporation of data, knowledge and feedback generated at the clinic into the basic research environment, and vice versa, back and forward.

Bioinformatics, and related fields within computational biology, contributes to such objectives with methodologies and technologies that facilitate a better understanding of biological systems and the connections between health and disease. As shown in the next chapters, this requires the analysis, visualization, modelling and integration of different types of data. It should be evident that this has nothing to do with 'number crunching' exercises or information technology service support. Bioinformatics is at the centre of an iterative, incremental process of questioning, engineering and discovery. This in turn allows researchers to improve their knowledge of the subtle relation between health and disease, and gives way to a capacity to predict events rather than simply describe them. Bioinformatics then becomes a translational discipline, that is 'translational bioinformatics', a major player in the development of a more predictive, personalized medicine.

Hypotheses about biological function and disease are typically made at the 'wet laboratory'. However, in a translational biomedical context, it is at the 'bedside' where medically-relevant questions and requirements may be initially proposed and where biological samples (fluids and solid tissue) are acquired from patients. This, together with a diverse range of data about clinical responses and life-styles, provides the inputs to different information platforms and processes. The resulting biological samples are processed in the laboratory to extract different types of molecular data, such as DNA sequences and the expression of genes and proteins. These questions and information are expanded, redefined and explored by biologists and bioinformaticians in close cooperation with clinical researchers.

Computational approaches and resources are required at both the clinic and the laboratory. This is not only because informatic infrastructures and large-scale data analysis are routinely required in these environments, but also because bioinformatics can directly specify and address questions of scientific and clinical relevance. In the post-genome era, this requires provision of alternative views of phenomena that goes beyond the single-gene, hypothesis-driven paradigm. Figure 1.1 illustrates examples of key aspects in the dialogue between the clinical, laboratory and computational research.

Within biomedical translational research, bioinformatics is crucial for accomplishing a variety of specific challenges: From the implementation of laboratory management systems, drug target discovery, through the development of platforms for supporting clinical trials, to drug design. This book will focus on computational and statistical approaches to disease biomarker discovery. This includes the detection of disease in symptomatic and asymptomatic patients, the prediction of responses to therapeutic interventions and the risk stratification of patients.

## 1.2 Biomarkers: fundamental definitions and research principles

A biomarker is 'a characteristic that is objectively measured and evaluated as an indicator of normal biological processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention' (Biomarkers Definitions Working Group, 2001). According to this definition, biomarkers can be divided into three main types: 'Type 0' represents biomarkers used to estimate the emergence or development of a disease; 'Type 1'

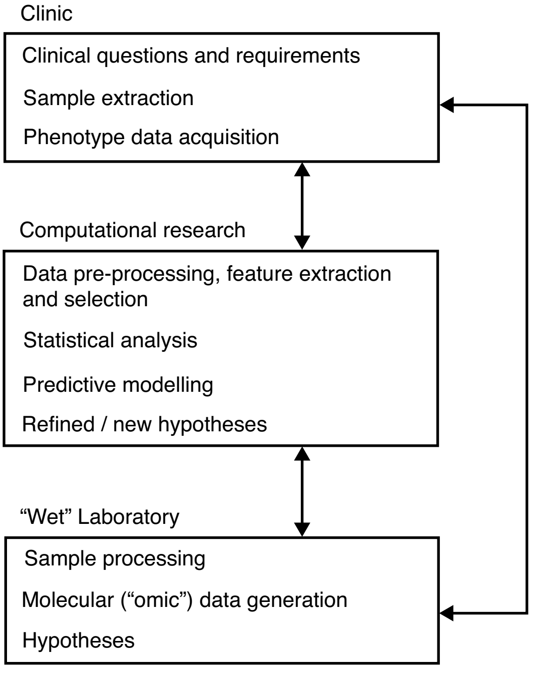


**Figure 1.1** The dialogue between clinical, laboratory and computational research environments in the context of translational biomedical research. Examples of typical tasks and applications

includes biomarkers that predict the responses to therapeutic interventions; and 'Type 2' represents biomarkers that, in principle, could be used as surrogate clinical endpoints in the course of clinical trials. An alternative classification that is commonly used in cancer research specifies two main types of biomarkers: predictive and prognostic biomarkers (Simon, 2008). The former refers to biomarkers used to predict therapeutic responses, and the latter refers to biomarkers for disease classification or risk estimation. This book will follow the categorization proposed by the US NIH Biomarkers Definitions Working Group.

On the basis of their application to the detection of disease, three main classes of biomarkers may be specified: screening, diagnostic and prognostic biomarkers. Screening biomarkers are used to predict the potential occurrence of a disease in asymptomatic patients. Diagnostic biomarkers are used to make predictions on patients suspected of having the disease. Prognostic biomarkers are applied to predict the outcome of a patient suffering from a disease. Most of the advances reported to date in the literature refer to diagnostic and prognostic biomarkers. This may be partly explained by the challenges posed by screening studies regarding the definition of complex phenotypes, independent evaluations and reproducibility of findings, and the lack of evidence showing their advantage in comparison with traditional disease risk factors.

Biomarkers can also be seen as indicators of functional and structural changes in organs and cells. Such changes may be associated with either causal factors (disease drivers) or consequences of normal and pathological events. Thus, biomarkers can be used to predict and monitor molecular changes relevant to the current development or future emergence of diseases, complications or responses. Moreover, biomarkers can also be considered as potential therapeutic targets, for example when their causal role in disease is demonstrated.

Clinical tests based on biomarkers have been applied for more than fifty years, but their potential applications for disease detection, patient stratification and drug discovery has expanded since the beginning of the twenty-first century. More recently, the discovery of novel biomarkers using genome-scale and different types of 'omic' data has become a crucial goal in both academia and industry. This interest has been driven in part by biomarkers' potential to predict disease states. Also biomarkers can facilitate a more comprehensive and deeper understanding of biological systems in the context of health and disease. Moreover, biomarkers can be used to guide the development of new therapies. For example, it has been suggested that biomarkers may reduce the time and costs of phase I and II clinical trials. This may be possible thanks to their potential as clinical endpoint substitutes (or surrogate endpoints), which are needed for assessing treatment safety and effectiveness.

The discovery of biomarkers is based on the following research principle: The comparison of physiological states, phenotypes or changes across control and case (disease) patient groups (Vasan, 2006; Gerszten and Wang, 2008). At the molecular level, such differences can be reflected in the differential activity or concentrations of genes proteins, metabolites and signalling pathways. Thus, biomarker discovery typically relies on the idea that those molecular species (i.e. gene, proteins, etc.) that display the greatest changes across phenotypes may be reported as potential biomarkers.

A traditional approach to discovering biomarkers for screening, diagnostic or prognostic purposes consists of the analysis of a single gene or protein and the identification of its 'abnormal' values, based on hypotheses biased toward specific biological processes or pathways. In general there are three traditional methods for identifying abnormal biomarker values: identification based on reference thresholds, based on discrimination thresholds and based on risk thresholds (Vasan, 2006). In the first approach the distribution of biomarker values in a reference group that approximates the general population is estimated and abnormal values are defined using extreme values on the basis of percentile thresholds. For example, a protein concentration value above the 99th percentile value can be considered abnormal and an indication of disease or clinical outcome. Discrimination thresholds can be defined after comparing the distribution of biomarker values between patient groups (e.g. control vs. disease) in terms of their differences or overlaps. For instance, a protein concentration value greater than 100 pg/mL may be associated with a specific clinical complication or disease. A discrimination threshold would aim to maximize the capacity of distinguishing between these groups. The approach based on risk thresholds aims to detect biomarker values that would be associated with a (disease or response) risk increase beyond a critical point on follow-up. For example, a systolic blood pressure value below 115 mmHg may be defined as 'desirable', as a value above this limit is linked to an increase of the risk of vascular disease.

Independently of their categorization, application domain or discovery approach, a fundamental objective in biomarker research is to detect a disease, response or complication at an early stage to aid in the selection of a treatment strategy. Such a prediction process should be sufficiently non-invasive, reproducible and inexpensive. In some clinical areas another important quality criterion is to maximize the predictive specificity (or reduction of false-positive rates, for example low rate of control patients incorrectly assigned to a pathological condition). This optimization is important because even relatively small false-positive rates can lead to unnecessary and expensive diagnostic or treatment procedures. In other areas the cost of missing a potential 'true positive' prediction of the disease is the main priority. Therefore, the selection and interpretation of prediction quality indicators are domain-specific, and may require the combination and optimization of different clinically-meaningful indicators. Chapters 2 and 3 include more detailed discussions on the evaluation of biomarkers and prediction models.

These prediction tasks will also directly influence the capacity to offer a more personalized management and treatment of patients. Moreover, it has applications in the assessment of therapeutic efficacy and toxicity. These prediction models can aid in the selection of those patients for whom treatment could offer an optimal benefit, and which could in turn reduce unnecessary therapy on patients with a better expected clinical outcome. Overall, this may directly contribute to the reduction of treatment and hospitalization costs.

## 1.3 Clinical resources for biomarker studies

Biomarker research relies on two main types of data acquisition strategies (Pepe *et al.*, 2001): Retrospective and prospective studies.

*Retrospective studies.* These studies are based on clinical samples collected before the design of the biomarker study, and before any comparison with control samples have been carried out. After a pre-determined period of follow-up, clinical outcomes or phenotypes are specified, and case and control samples are compared. Biomarker discovery based on retrospective studies looks back at past, recorded data to find evidence of marker-disease relationships.

Depending on the study objectives, the control samples may be derived from healthy populations or from those subjects that did not show the positive clinical outcome under study (e.g. individuals who did not develop the disease, die or show complications). These studies may involve the identification of biomarkers to distinguish between patients at first time of consultation, or as a function of time (i.e. several clinical evaluation times) before determining the predictive capacity of the biomarkers. These studies also require investigations on the classification ability of covariates (other predictive cofactors), for example standard biomarkers or life-style information. Comparisons of multiple combinations of potential biomarkers with traditional biomarkers are fundamental. There is no universal standard for defining the length of follow-up times, which will be specific to clinical purposes, resource and biological constraints and economic costs. Matching of case-control samples on the basis of individual-based characteristics is important, as well as matching of subjects on the date of study enrolment when possible. Different classification quality indicators and techniques may be used to estimate the predictive or classification capacity of the biomarkers (Chapters 2 and 3). For instance, different prediction quality indicators, corresponding to the different follow-up periods can be estimated and compared for different classification models. The main goal is to identify those prediction models capable of identifying patients with the clinical outcome at a number of months (or years) after the biomarkers are measured.

*Prospective studies*. In this type of study, the biomarker-based prediction or classification model is applied on patients at the time of patient enrolment. Clinical outcomes or disease occurrence are unknown at the time of enrolment. Thus, selected subjects are followed during a pre-determined period time, that is prospective studies look forward in time. At the end of such a period, information about the clinical outcomes is acquired and analysed to assess the prediction or discrimination capacity of the biomarkers.

In some applications, such as the independent validation of a new biomarker model in a real clinical setting, those patients testing positive would undergo further diagnostic or prognostic procedures. This will allow the estimation of the model capacity to detect true positive cases, disease stage and other characteristics. In addition, these studies would not only drive the classification or risk assessment of patients, but also the selection of treatments.

In a biomarker development project, prospective studies typically follow the completion of retrospective studies in order to further evaluate the clinical potential of the proposed biomarkers and prediction models. Although more expensive and time-consuming, prospective studies are considered a less biased and more objective approach to collecting and analysing data for biomarker discovery.

## 1.4 Molecular biology data sources for biomarker research

Traditional and large-scale molecular biology generates data needed to reflect physiological states in modern biomarker discovery. The availability of new data sources originating from different 'omic' approaches, such as genomic variation and mRNA expression analysis, are allowing a more systematic and less biased discovery of novel biomarkers in different clinical areas. Moreover, some of such new biomarkers are orthogonal, that is biomarkers with relatively low statistical, biological or clinical dependencies between them.

Major sources of molecular data for biomarker discovery are (Vasan, 2006; Gerszten and Wang, 2008): DNA-based variation studies (Chapter 4), gene expression or transcriptomics (Chapter 5), protein expression and large-scale proteomics, and the measurement of metabolite and small molecule concentrations (metabolomics) (Chapter 6).

In genomic variability studies, a key discovery approach is the analysis of single-nucleotide polymorphisms (SNPs) in cases versus control subjects. Variants with potential screening, prognostic or diagnostic potential have been proposed based on the analysis of candidate genes and genome-wide association studies (Chapter 4) in different medical areas, including cancer and cardiovascular research. However, the independent validation or reproducibility of these results has been proven to be more difficult than anticipated. Examples of recent advances include SNPs biomarkers for early-onset of myocardial infarction and premature atherosclerosis (Gerszten and Wang, 2008).

In some areas, such as cardiovascular research, the discovery of disease biomarkers using gene expression analysis has been traditionally limited by the difficulty in obtaining tissue samples. Different studies using cardiomyocites in culture, *in vitro* models and tissue extracted from transplant patients have suggested a great variety of potential diagnostic and prognostic biomarkers, for example mortality in patients with heart failure. The development of less invasive techniques based on peripheral blood gene expression profiling represents a promising approach in this and other medical domains (Chapter 5).

Proteomics and metabolomics have become promising technologies for biomarker discovery. These technologies enable the analysis of the clinically-relevant catalogues of proteins and metabolites (Chapter 6). Metabolites are sets of biochemical substances produced by metabolic processes (e.g. sugars, lipids and amino acids). These approaches represent powerful complementary views of the molecular state of a cell at a particular time. A major challenge is the diversity of cell types contributing to the human proteome and metabolome (e.g. plasma proteome) and the low concentration levels of many of the proteins suggested as disease biomarkers. On the other hand, it has been suggested that the size of the human metabolome might be represented by a relatively small set ($\sim$3000) of metabolites (Gerszten and Wang, 2008).

Independently of the types of 'omic' resources investigated, there is the possibility that the molecular profiles or patterns observed in the potential biomarkers may not be true reflections of primary molecular events initiating or modulating a disease. Instead, they may reflect a consequence of downstream events indirectly caused by the studied pathology at later stages.

Modern biomarker discovery research aims to extract information from these resources, independently or in an integrated fashion, to design predictive models of disease occurrence or treatment responses. The integration of different types of clinical and 'omic' data also motivates the extraction of biological knowledge from diverse distributed repositories of functional annotations and curated molecular pathway information (Ginsburg, Seo and Frazier, 2006; Deschamps and Spinale, 2006; Camargo and Azuaje, 2007) (Chapter 7). This, in turn, promotes the implementation of advanced predictive integration-based approaches (Chapter 8), that is biomarker-based models of disease or treatment response that combine quantitative evidence extracted from different data sources (Camargo and Azuaje, 2008; Ideker and Sharan, 2008). These tasks are facilitated through significant computational advances accumulated over the past 20 years in connection with information standardization, ontologies for supporting knowledge representation and exchange, and data mining (Chapter 9).

## 1.5 Basic computational approaches to biomarker discovery: key applications and challenges

Advances in computational research and bioinformatics are essential to the management and understanding of data for biomarker discovery. Examples of such contributions are the storage (including acquisition and encoding), tracking (including laboratory management systems) and integration of data and information (Azuaje, Devaux and Wagner, 2009a, 2009b). Data integration involves the design of 'one-stop' software solutions for accessing and sharing data using either data warehousing or federated architectures. This has allowed a more standardized, automated exploration, analysis and visualization of clinical and 'omic' data using a great variety of classic statistical techniques and machine learning (Azuaje, Devaux and Wagner, 2009a, 2009b).

Biomarker discovery from 'omic' data also relies on exploratory visualization tools, data clustering, regression and supervised classification techniques (Frank *et al.*, 2004;

Camargo and Azuaje, 2008). Feature selection (Saeys, Inza and Larrañaga, 2007) also represents a powerful approach to biomarker discovery by exploiting traditional statistical filtering (e.g. statistical analysis of multiple hypotheses) or models ‘wrapped’ around classifiers to identify powerful discriminators of health and disease (Chapter 3). The resulting significant features can then be used as inputs to different machine learning models for patient classification or risk estimation, such as neural networks and support vector machines (Chapter 3).

Other important challenges for bioinformatics are the relative lack of data together with the presence of different potential sources of false positive biomarker predictions, including experimental artefacts or biological noise, data incompleteness and scientific bias (Ginsburg, Seo and Frazier, 2006; Azuaje and Dopazo, 2005; Jafari and Azuaje, 2006). This further adds complexity to the task of evaluating the predictive capability of disease prediction models, particularly those based on the integration of multiple biomarkers.

A key challenge in biomarker development is the reduction of experimental variability and noise in the data, as well as the accomplishment of reproducibility at the different stages of sample acquisition, measurement, data analysis and evaluation. Potential sources of experimental variability are related to sample extraction, data storage and processing. This may result in inter-laboratory variability driven by factors such as diversity of reagents, experimental platforms and protocols. Recommendations and standards have been proposed by technology manufactures and international community groups, which define practices for sample handling, quality control and replication.

Apart from minimizing variability related to experimental factors, it is crucial to address patient- and data-related sources of variability, such as intra- and inter-individual variability. Such variability may be caused by factors ranging from age, gender and race to drug treatments, diet or physical activity status. Depending on the suspected factors influencing these differences, prediction model stratification or statistical adjustments may be required. Standards and recommendations for supporting better reproducibility of data acquisition (e.g. MIAME) and analysis (e.g. replicate and pre-processing procedures) have also been proposed by manufacturers and the international research community (Brazma, Krestyaninova and Sarkans, 2006). Additionally, the accurate and sufficient reporting of biomarker studies, for example diagnostic accuracy results, has motivated the development of specific community-driven guidelines (Chapter 10).

Research in bioinformatics shares the responsibility to lead efforts to standardize and report biomarker study results, to provide extensive prediction model evaluation, and to develop advanced infrastructures to support research beyond the ‘single-marker’ analysis approach. There is still a need to develop more user-friendly tools tailored to biomarker discovery, which should also be able to operate in open and dynamic data and user environments. Despite the availability of ‘generic’ bioinformatic tools, such as statistical analysis packages and platforms for the design of machine learning systems, the biomarker research community will continue requiring novel solutions to deal with the requirements and constraints imposed by the translational research area. Table 1.1 reflects the diversity of computational technologies and applications for biomarker discovery. It shows how different requirements and problems are connected to specific fields and technologies.

**Table 1.1** Examples of key computational technologies and applications for biomarker discovery. Circles inserted in cells represent a significant connection between a bioinformatics technology or research area (columns) and applications relevant to biomarker discovery research (columns)

| | Stat | ML | GNT | IV | KE | SD | SM |
|---|---|---|---|---|---|---|---|
| Estimation of significant relationships/differences between patients | • | • | | • | | | |
| Selection of optimum biomarker sets | • | • | • | | | | |
| Integrated access to data and information | | | | | • | • | |
| Integrated analysis of data and information for prediction modelling | | • | • | | | | • |
| Laboratory information management and tracking systems | | | | | • | • | |
| Biobanks | | | | | • | • | |
| Literature search and mining | | | | | • | | |
| Data and information annotation | | | • | | • | • | |
| Discovery infrastructures, automated distributed services | | | | | • | • | |
| Patient classification and risk score assessment | • | • | | | | | |

Stat: Statistical analysis including hypothesis testing, ML: Statistical and machine learning, GNT: Graph and network theory, IV: Information visualization, KE: Information and knowledge engineering and management, including natural language processing, SD: Software development and Internet technologies, SM: Complex systems modelling including simulation tools.

From a data analysis perspective, biomarker discovery can be seen as an iterative, incremental process (Figure 1.2). Differential pattern recognition, classification, association discovery and their integration with diverse information resources, such as functional pathways, are central to this idea. The main expected outcomes, from a translational research perspective, could be new diagnostic or prognostic kits (e.g. new biochips or assays) and computational prediction systems for screening, diagnostic and prognostic purposes. The validity and potential clinical relevance of these outcomes will depend on the successful implementation of evaluations using independent samples. Moreover, the applicability of new biomarkers, especially multi-biomarker prediction models, will also depend on their capacity to outperform conventional (or standard) markers already incorporated into the clinical practice.

Bioinformatics research for biomarker discovery also exploits existing public data and information repositories, which have been mainly the products of several publicly-funded initiatives. Different approaches have shown how novel biomarker discovery based on the integrative data analysis of different public data sets can outperform single-resource (or single site) studies, and provide new insights into patient classification and

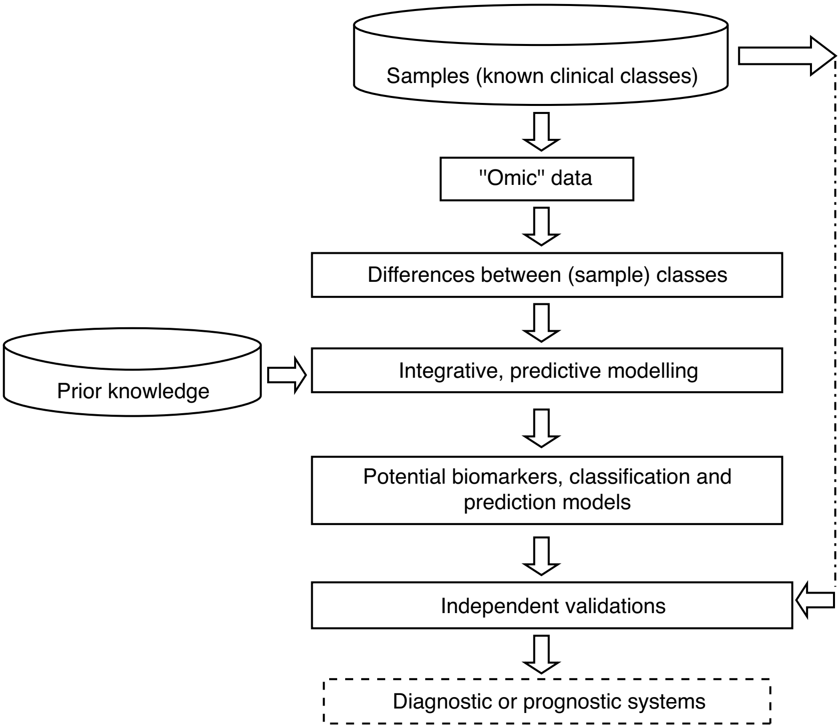


**Figure 1.2** A typical biomarker discovery framework

processes underlying a disease (Camargo and Azuaje, 2008; Butte, 2008a, 2008b). Currently different online projects and repositories freely offer genomic variation data (e.g. human genomic polymorphisms), gene expression (e.g. public repositories with raw public data), proteomics (e.g. plasma proteome, antibodies), and human health-specific pathways (e.g. metabolic, signalling and genetic interactions) (Mathew *et al.*, 2007). Chapter 9 will review relevant bioinformatic infrastructures, information resources and software tools for supporting biomarker discovery. Chapters 8 and 10 will discuss the analysis of multiple public data and information resources.

## 1.6 Examples of biomarkers and applications

Molecular biomarkers are measured in biological samples: solid tissues, blood or other fluids. In the area of cardiovascular diseases, for example, a typical clinical situation for the application of biomarkers is when a patient presents severe chest pain. This would trigger questions such as: Is this patient experiencing a myocardial infarction or unstable angina? If the patient is experiencing a myocardial infarction, what is the likelihood that this patient will respond to a specific therapy? What is the amount of myocardial damage? What is the likelihood of a future recurrence? What is the likelihood of progressing to heart failure or death in the near future? Protein biomarkers, for instance, may be applied to help doctors to answer these questions.

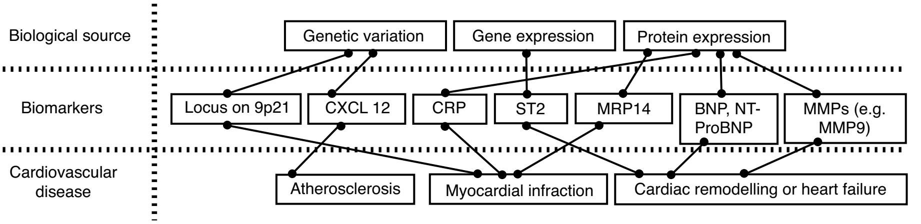


**Figure 1.3** Examples of cardiovascular biomarkers and their relationship to different 'omic' technologies and diseases (Vasan, 2006; Gerszten and Wang, 2008)

In principle, new biomarkers will be of clinical value only if the following factors can be demonstrated: predictive or classification accuracy, reproducibility, their acceptance by patient and clinician, high sensitivity and specificity, direct relation with changes in a pathology or clinical conditions, and measurable impacts in patient management. However, depending on the type of application, some of these (and other) factors will be more or less relevant. In screening applications, high predictive performance quality (e.g. overall accuracy, sensitivity and specificity) and relative low costs could be the most critical factors. These factors are also important in diagnostic applications of biomarkers, together with other factors, such as high tissue specificity and potential to be applied at point-of-care setting. In some prognostic applications, quality indicators such as specificity and sensitivity may be less critical than the reduction of intra-individual variation. Chapter 10 provides a more detailed discussion on the assessment of clinical relevance in biomarker research.

The increasing availability of large-scale data sources originating from diverse 'omic' approaches, such as genomics and transcriptomics, are allowing a more systematic and less biased discovery of novel disease biomarkers in different clinical areas. Figure 1.3 illustrates relevant examples of biomarkers from the cardiovascular research area, which are based on different types of 'omic' approaches and technologies.

Examples of diagnostic cardiovascular biomarkers incorporated into clinical practice are the brain natriuretic peptide (BNP) for heart failure, and troponin I and troponin T for myocardial infarction (Gerszten and Wang, 2008). In addition, it has been suggested that these biomarkers also have prognostic applications. Examples of potential screening biomarkers include those that may be associated with inflammation (e.g. C-reactive protein and interleukin-6), thrombosis (e.g. fibrinogen) and other vascular complications (Gerszten and Wang, 2008). However, it is important to stress that their clinical utility still remains a topic of exploration and discussion. The capacity of novel prediction or classification models, based on the combination of novel biomarkers, to outperform traditional biomarkers has not been widely demonstrated. For instance, a report from the Framingham Heart Study evaluated the predictive capacity of several molecular markers of death and cardiovascular complications. This investigation concluded that multi-marker prediction models can only add a moderate improvement in prediction performance, in comparison with (single-marker) conventional models. However, these relative small effects may also account for an over-emphasis put on standard quality indicators for sample classification without adequately considering other measures and design factors, such as specific prediction goals and sample class imbalances (Chapter 3).

Another potential obstacle is that the majority of reported biomarkers may be biased towards well-studied functional pathways, such as those linked to inflammation and cholesterol synthesis in the case of cardiovascular biomarkers (Gerszten and Wang, 2008). Moreover, multi-marker models may be based on correlated biomarkers, which may in turn reduce the classification ability of the models. In the data mining and machine learning areas it is well-known that, for classification purposes, the combination of several correlated predictive features is less informative than the combination of fewer uncorrelated (or orthogonal) biomarkers. These difficulties and limitations are also found in other medical areas.

Natriuretic peptides have become the major prognostic reference in heart failure diagnosis, prognosis and treatment planning (Maisel, 2007). In particular, BNP and NT-proBNP have become powerful indicators of heart failure in acute dyspnoea patients and of clinical outcomes in advanced heart failure. The correlation between their patho-physiology and heart failure is strong enough to allow, for example, effective treatment of some patients through the exogenous administration of BNP. Thus, this is an example of a biomarker that satisfies some of the key requirements in biomarker discovery: biomarkers should not only represent strong indicators of disease, but also they should be useful for the early detection and treatment of the disease. BNP levels have also been used to indicate admission in emergency units, level and types of treatments, as well as prognosis during treatment. For example, low BNP levels in patients under treatment may call for the application of additional treatments (Maisel, 2007).

However, natriuretic peptides have not been widely adopted for robust, accurate patient stratification or for the early detection of heart failure onset. For example, strong correlations between some levels of BNP (e.g. 100–400 pg/ml) and clinical outcomes may not always be possible to observe, and there are important level overlaps between different clinical groups (Maisel, 2007). Moreover, for patient classification or screening, there is no conclusive evidence on how this information may consistently be applied to improve classification sensitivity or specificity in comparison to more traditional methods. This is another reason to explore the potential of multiple biomarkers integrated by advanced statistical analysis and machine learning techniques.

Recent advances in the use of multiple biomarkers include the prediction of death from cardiovascular disease in the elderly (Zethelius *et al.*, 2008). In this example, protein expression biomarkers relevant to different functional pathways, such as cell damage and inflammation, improved risk prediction in comparison to traditional clinical and molecular biomarkers, such as age, blood pressure and cholesterol. The proposed and reference prediction models were based on traditional survival analysis during a follow-up period of more than 10 years, and were comparatively evaluated using standard indicators of predictive quality (Chapters 2 and 3).

## 1.7 What is next?

The next chapters will discuss the analysis of different types of 'omic' data for identifying and evaluating disease biomarkers, including diagnostic and prognostic systems. It will offer principles and methods for assessing the bioinformatics/biostatistics limitations, strengths and challenges in biomarker discovery studies. Examples of studies and applications based on different techniques and in several clinical areas will be explained.

Descriptions and discussions take into account the diverse technical backgrounds and research roles of the target readership. A major objective is to increase the awareness of problems, advances and possible solutions. But also we expect that it will facilitate a more critical approach to designing, interpreting and evaluating biomarker studies. The book targets users and designers of bioinformatic technologies and applications. Similarly, it can benefit biologists and clinical researchers with an interest in improving their knowledge of how bioinformatics can contribute to biomarker discovery. Readers will benefit by learning about: (1) key requirements and diversity of data resources available for biomarker discovery in different clinical domains; (2) statistical and data mining foundations for identifying, selecting and evaluating biomarkers and prediction systems; (3) major advances and challenges in bioinformatics and biomarker research; (4) computational and statistical requirements for implementing studies involving different types of biomarkers; (5) major bioinformatic advances and approaches to support biomedical translational research; and (6) the integration of 'omic' data and prior knowledge for designing and interpreting prediction models.

Although the book will emphasize examples of problems and applications in cardiovascular and cancer research, the computational solutions and advances discussed here are also relevant and applicable to other biomedical areas. Some of the chapters will be complemented by short commentaries from highly esteemed researchers to provide alternative views of biomedical problems, technologies and applications.

This book will focus on how fundamental statistical and data mining approaches can support biomarker discovery and evaluation. Another key aspect will be the discussion of design factors and requirements for implementing emerging approaches and applications. The book will not deal with specific design or implementation problems related to pharmaceutical research and development, such as the assessment of treatment responses in drug clinical trials. However, many of the design and evaluation techniques covered here may be extended to different problems and applications.

The next two chapters are 'foundation' chapters, which will provide readers with the knowledge needed to assess the requirements, design tasks and outputs of disease biomarker research. These sections also introduce some of the most relevant computational approaches and techniques for 'omic' data analysis. This will be followed by detailed discussions of methodologies and applications based on specific types of 'omic' data, as well as their integration for biomarker discovery. Such chapters will reflect the 'how' and 'what' aspects of these research areas. Chapters 9 and 10 will focus on the critical assessment of key bioinformatic resources, knowledge gaps, and challenges, as well as emerging and promising research directions. These final sections will underscore the 'why' and 'when' aspects of problems and applications. Thus, one of the main goals is to focus on fundamental problems, common challenges across information types and clinical areas, and design principles.

At this point, it is necessary to introduce basic mathematical notation and terminology to facilitate the understanding of the techniques and applications. For most statistical and machine learning analyses, it will be assumed that data sources can be, at least to some extent, represented as data matrices. Capital letters in bold will be used to refer to this type of resources. For example, $\mathbf{D}$ represents a data set with $m \times n$ values, with $m$ representing the number of rows, and $n$ representing the number of columns in $\mathbf{D}$. A row (or column) can represent samples, biomarkers or other 'omic' profiles, which will be represented by bold and lower case letters. For example, $\mathbf{s}$ represents a vector of $m$ values, with biomarker values extracted from a single patient. References to individual data values will be expressed by using lower case letters. Subscripts will be used to refer to specific vectors (e.g. samples or biomarkers) or values. The term ‘class’ will be used to refer to specific phenotypes, patient groups or biological processes. When using networks to represent data, a network node will represent a biological entity, such as a gene or potential biomarker. A network edge, linking two or more nodes, encodes any biologically-meaningful relation, such as different types of functional interactions.

It is evident that time and publication space constraints would not allow one to cover all major methodologies, tools and applications in detail. However, the content of the chapters have been selected to avoid, or at least reduce, methodological bias or preferences for specific data mining techniques or algorithms. This is particularly relevant when one considers the speed of progress in computational and data analysis research. Therefore, the book structure has been shaped by major (‘omic’) data types and problems, rather than specific techniques.

Although a spectrum of data mining techniques for biomarker discovery will be introduced in Chapter 3, the book does not intend to offer a detailed coverage of specific algorithms or techniques. Emphasis will be put on design and evaluation requirements and questions, interpretation of inputs and outcomes, adaptation and combination of approaches, and advanced approaches to combining hypothesis- and discovery-driven research.

# 2 Review of fundamental statistical concepts

This chapter offers a brief introduction to basic statistical knowledge for helping the reader to design and interpret disease biomarker discovery studies: Statistical error types, data sampling and hypothesis testing for numerical and nominal data, odd scores, and the interpretation of other statistical indicators. This includes parametric and non-parametric techniques for comparing groups and models, which is a basic approach to detecting potential biomarkers in large-scale 'omic' studies. This chapter also explains different predictive evaluation techniques: traditional measures, techniques for classification and numerical predictions, and an introduction to the application of receiver operating characteristic curves and related methods.

## 2.1 Basic concepts and problems

Although we assume that the reader has some basic experience in statistical analysis, the first half of this chapter offers a quick overview of fundamental definitions and terminology that will be used in subsequent chapters.

A key approach to biomarker discovery research is to compare cases *vs.* control samples to detect statistical differences, which could lead to the identification and prioritization of potential biomarkers. Control and case samples are commonly obtained before treatment or before knowing its classification (e.g. diagnosis, prognosis). Control samples are obtained from healthy patients, untreated patients, or from patients who did not experience the specific clinical outcome under analysis. The pairing or matching of control and case patients is a strategy to prevent irrelevant factors to confound the observed associations or predictions. In this scheme the control and case groups are formed by selecting pairs of samples (e.g. patients) sharing common characteristics that

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

may represent confounding factors. Irrelevant or confounding factors may refer to molecular, clinical or environmental variables that are not directly related to the disease. In relatively small groups of samples, careful sample matching between the compared groups is recommended instead of random samplings of the populations. The random assignment of individuals to small case and control groups may generate statistically detectable differences on the basis of factors such as age, sex, and different life-style factors. Chapter 4 offers a more detailed discussion on population stratification and confounding factors.

Important design factors to be considered in the implementation of biomarker discovery also include the potential confounding effects of drug treatments on the population under study, the relative lack of data, and the presence of spurious findings due to insufficient statistical evaluation. Multi-stage studies are recommended to improve the quality of potentially significant predictions and to reduce costs (Chapter 4). Thus, investigations carried out on relatively small sample groups from carefully selected cohorts typically precede larger and more heterogeneous studies.

In biomarker discovery, researchers regularly need to characterize data on the basis of patient groups, pathologies, clinical responses and molecular function. Different statistical descriptors can be used to summarize such characteristics and differences. Such descriptive statistics are later used to make and test hypotheses about the patient populations and potential biomarker sets under study.

The properties of interest, for example molecule concentrations and values of clinical risk factors, can be represented by either discrete or continuous (numerical) values and are commonly referred to as variables or features. Discrete features can represent nominal and ordinal data. The former refers to categories with no particular order, such as gender, yes/no values. The latter refers to categories that reflect some meaningful ordering, such as prognostic classes encoded by disease grades or low-medium-high labels. The selection of descriptive statistics, hypothesis testing procedure and prediction models depend on the type of feature investigated.

Discrete or categorical data are typically summarized by using frequency descriptors: absolute, relative or cumulative. Continuous features can be represented by measures of centrality (e.g. mean, median, mode), and dispersion measures (e.g. standard deviation, variance). Dispersion measures describe the variation of the data around measures of centrality. For more detailed descriptions of these and related measures, as well as of basic data display techniques, the reader is referred to (Glantz, 2001) or (Larson, 2006).

Based on descriptive statistics one can estimate properties that are representative of a population. The main goal of estimation from data is to approximate such properties, such that they can be seen as representative of a general population, for example the population of patients with a disease, or the population of patients that respond positively to a drug. Typically, researchers make point or single-value estimates (e.g. the mean) and confidence interval (CI) estimates. A CI represents a range of values around a point estimate with a specific level of confidence. In this case researchers typically report the 95% CI that would include the true value of the statistical parameter being estimated. When analyzing relatively large data sets, the 95% CI is almost twice the standard error (SE) of the point estimate (1.96 times SE). For instance, at the 95% CI one can say that the true value of the mean in a population is above (or below) the mean value estimated from the data samples by an amount equal to 1.96 times SE. The SE can also be estimated from

**Table 2.1** Estimation of means, proportions and confidence interval (CI) for typical comparison scenarios and different data types in biomarker research

| Data type | Estimation scenario | Estimated value | CI |
|---|---|---|---|
| Numerical, continuous | Single population | Mean | $M \pm v \times SE(M)$ |
| Numerical, continuous | Two independent populations | Mean differences | $(M_1 - M_2) \pm v \times SE(M_1 - M_2)$ |
| Numerical, continuous | Paired populations | Mean differences | $M_d \pm v \times SE(Md)$ |
| Discrete | Single population, two categories | Proportion differences | $Pt \pm v \times SE(Pt)$ |
| Discrete | Two independent populations | Proportion differences | $(Pt_1 - Pt_2) \pm v \times SE(Pt_1 - Pt_2)$ |

*M*: estimated mean, *Pt*: estimated proportion of category studied. Subscript indices are added to *M* and *Pt* to represent different groups. *SE*: standard error of the mean, mean difference or proportion analysed (between parentheses), *v*: statistic value from the data distribution (e.g. the *t* distribution) associated with the required confidence level. In the example presented above $v = 1.96$ for a 95% confidence. When using discrete data, *v* can be obtained from the *Z* distribution, for example. $M_d$: mean difference in paired data. Additional information, including the calculation of *v* and *SE* values, can be obtained in (Glantz, 2001; Sullivan, 2006).

the data available. Where does the 1.96 come from? This number and $-1.96$ are the values at which 95% of the area under the standard normal distribution is obtained. Thus, different confidence levels will produce different confidence intervals around the parameter estimated.

In biomarker discovery, estimates are commonly needed for analyzing a single or two (or more) populations. A typical application in the latter scenario involves control and case population samples. A case class may represent a specific disease or treatment response. Moreover, it is necessary to specify whether such populations are independent (i.e. different patients from different groups), or paired populations (e.g. each patient provides samples under different conditions or at different times). Answers to these questions will guide the selection of estimation and hypothesis testing methods. Table 2.1 summarizes how to compute point and CI estimates for different types of data and typical comparison scenarios, using means and proportions. For more detailed discussions, the reader may refer to (Glantz, 2001) or (Sullivan, 2006).

Estimates of central points and dispersion are used to compute statistical scores for different hypothesis testing applications (next section), for example the *t* statistic for comparing means in a microarray analysis and the $\chi^2$ statistic for comparing proportions of phenotype categories in genotype-phenotype association studies. These statistic values are used to estimate probability, *P*, values that can be interpreted to reject or accept a hypothesis. In this case, *P* is an estimate of the probability that the observed estimates (or differences) are statistically detectable at a particular significance level ($\alpha$). To put it another way, *P* is an estimate of the probability of observing, by chance, a statistic value (e.g. *t* statistic) as large (or larger) as the observed statistic value.

These *P* values are obtained from probability tables that are tabulated for different statistical distributions, degrees of freedom (df) and test directionality. The df are determined by the number of groups compared and the number of samples. The terms

**Table 2.2** Definition of error types in the context of a typical hypothesis testing study, such as the comparison of a control and experimental group on the basis of a biomarker concentration value. TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

| | Null hypothesis is true | Null hypothesis is false |
|---|---|---|
| Null hypothesis rejected | Type I error (FP) | TP (correct prediction) |
| Null hypothesis not rejected | TN (correct prediction) | Type II error (FN) |

one-tailed (or one-sided) and two-tailed (or two-sided) are used to define test directionality. A two-sided test means that the null hypothesis (see below) is rejected if the test statistic falls on either side (tails) of the data distribution considered. Estimates of *P* values also depend on several assumptions about the data distribution modelling the problem and the characteristics of the populations under study (e.g. assumption of equal variances between the sample groups).

A statistical 'significance' analysis requires the researcher to specify a null hypothesis and an alternative hypothesis before making any calculations. The null hypothesis, *Ho*, typically refers to the absence of effects or differences in the problem investigated. For example, there is no difference between healthy and disease group on the basis of age. The alternative hypothesis, *Ha*, is the hypothesis that the researcher aims to demonstrate. Thus, the *P* value from a hypothesis testing procedure estimates the probability of obtaining the observed statistic value under the null-hypothesis.

The correct or wrong rejection of the null hypothesis (or acceptance of the alternative hypothesis) defines Type I and II errors, which are commonly used to interpret the predictive quality of a particular hypothesis testing procedure or prediction algorithm. If the null hypothesis is incorrectly rejected then a false positive prediction is reported, that is a Type I error. If the null hypothesis is false and one fails to reject it, then a false negative prediction has been made, that is a Type II error. Table 2.2 illustrates these error types in the context of a typical hypothesis testing study, such as the comparison of a control and experimental group on the basis of a biomarker concentration value. Note that the interpretation of these errors can also be applied to classification problems, such as biomarker-based classification of patients belonging to two different diagnostic classes. Table 2.3 illustrates this scenario. In multiple-hypotheses testing applications, the false discovery rate is introduced by some procedures to estimate the proportion of incorrect rejections of null hypotheses (false positives) that the researcher is willing to consider amongst the rejected hypotheses.

**Table 2.3** Definition of prediction errors in a typical diagnostic (or classification) application. TP: True Positive, TN: True Negative, FP: False Positive, FN: False Negative

| | Disease absent | Disease present |
|---|---|---|
| Biomarker predicts presence of disease | FP | TP (correct prediction) |
| Biomarker predicts absence of disease | TN (correct prediction) | FN |

## 2.2 Hypothesis testing and group comparison

Hypothesis testing is the process of inferring conclusions about data based on the application of statistical tests. These procedures offer answers to questions such as: Are there any significant differences between control and case patients on the basis of a specific biomarker value? Does the mean age in this population significantly differ from 35 years? The outcomes of a statistical testing procedure are statistic and *P* values, which estimate the strength of a hypothesis (*Ha*) in relation to the null-hypothesis (*Ho*). This allows the researcher to make inferences about a population based on the dataset (sample) available.

As pointed out in the previous section, the researcher should first of all define specific, mutually exclusive, null and alternative hypothesis. The former is commonly defined to specify the lack of effects, differences or associations in a study. The latter specifies the characteristic or finding that the researcher aims to demonstrate. Consider the example in which a group of case subjects, **D**, with a disease is compared to a control group, **H**, on the basis of the concentration value of protein, protY. In this case the researcher may want to report potential significant differences between these patient groups to support the hypothesis that protY could be a diagnostic biomarker of the disease. If we use the means of the protein concentration values to describe each group, *Ho* can be expressed as: ‘there is no detectable difference between the groups, **D** and **H**, in terms of their respective mean protY concentration values’. On the other hand, *Ha* establishes that ‘there is a potentially important difference’.

As in the case of a jury trial (Davis and Mukamal, 2006), one can only reject *Ho*, or fail to reject it, based on the evidence available. But if the researcher cannot reject *Ho* (e.g. high *P* values suggesting no significant difference), this does not mean that *Ho* is necessarily true. As pointed out above, the researcher has to define a significance level in advance, which is usually equal to 5% or $P = 0.05$, to indicate the probability of (incorrectly) rejecting *Ho* when *Ho* is actually true (i.e. a false positive prediction).

The previous section revised the different types of errors that may occur when rejecting or failing to reject a true *Ho*. This now can be complemented with the concept of ‘power’ of the statistical test, which is the probability of correctly rejecting *Ho*. Thus, the power of a statistical test is defined as $(1 - \beta)$, where $\beta$ is the probability of making a false-negative prediction. One can also say that $\beta$ represents the probability of failing to reject *Ho* when *Ho* should actually be rejected (a Type II error). For instance, a study in which one fails to reject *Ho* when a strong association between the biomarker and the disease is actually present.

In summary, before applying a hypothesis testing procedure, the researcher should specify in advance: hypotheses (*Ho* and *Ha*), the level of significance, statistical assumptions, and test directionality (e.g. two-sided). The method of analysis selected depends on the research problem. Two key factors that need to be considered are the types of data and comparisons. For both discrete and numerical data, typical applications include one- and two-sample analysis, with the latter including tests for paired and independent samples. Table 2.4 provides an overview of test selection criteria for different studies and data types. Mathematical details of these and other tests are explained in (Glantz, 2001; Davis and Mukamal, 2006; Gauvreau, 2006), including different examples from biomedical research.

**Table 2.4** Examples of test selection criteria for different studies and data types (Glantz, 2001). In one-sample analysis, the mean (or proportions) observed in a single group of patients, for example, is compared to a reference value. An example of comparison of independent samples is the analysis of two groups of different patients in terms of a single biomarker concentration value. Paired samples are analyzed when, for example, differences in a biomarker value are measured in the same group of patients, before and after the application of a treatment

| Data type | Comparison type | Compared measure | Key tests |
|---|---|---|---|
| Numerical | 1-sample | Mean | $z$-test, $t$-test |
| Numerical | Independent samples | Mean | $t$-test |
| Numerical | Paired samples | Mean | Paired $t$-test |
| Nominal | 1-sample | Proportions | $z$-test for proportions |
| Nominal | Independent samples | Proportions | Chi-2 ($\chi^2$) |
| Nominal | Paired samples | Proportions | McNemar's test |
| Ordinal | Independent samples | Ranks | Mann-Whitney rank-sum test |
| Ordinal | Paired samples | Ranks | Wilcoxon signed-rank test |

## 2.3 Assessing statistical significance in multiple-hypotheses testing

The hypothesis testing procedures reviewed above are applied to problems when a single variable is estimated and compared against a reference or across sample groups. In a typical biomarker discovery study a key challenge is to detect potentially significant differential responses or behaviours, such as differential gene expression, in a large set of variables (i.e. potential biomarkers) across control and cases groups. The number of variables could range from hundreds to thousands in a typical 'omic' biomarker discovery investigation. This type of analysis is referred to as multiple-hypotheses testing. This addresses the problem of underestimating the $P$ values obtained from the hypothesis testing of each variable independently.

In a multiple-hypotheses testing procedure involving, for example, $n$ genes, one would test each gene independently for differences between two groups using a significance level, $\alpha$. If the resulting $P_i$ value, for gene $i$, is smaller than $\alpha$, one can reject the $Ho$, and argue that there is a detectable or 'significant' difference between the compared groups in terms of gene $i$. But note that every time that a test is independently performed on each gene one is also admitting that there is a possibility that an error will be made. For example, 5% of the time when $\alpha = 0.05$. Now, it is clear that the more genes one analyses, the greater the possibility of reporting false positive predictions. For example, if $\alpha = 0.05$ and $n = 100$, one would expect that at least five genes could be found to be significantly differentially expressed when actually they are not, that is when $Ho$ should not be rejected. Hence, if $n$ genes are tested in the same biomarker discovery investigation, one would expect that some of the genes detected as 'significant' will actually be false positives. Thus, corrections or adjustments have to be made to these values to reduce the possibility of finding spurious findings. The multiple-hypotheses testing problem can be addressed through the estimation of the family-wise error rate (FWER), and the false discovery rate (FDR).

Suppose that we aim to implement $g$ statistical tests on a set of $g$ potential biomarkers. Thus, one is testing $g$ potential null hypotheses, $Ho^i$, $i = 1, \ldots, g$. The FWER is defined as the probability of making false positive predictions (type I errors) amongst the $g$ hypotheses. Multiple-hypotheses testing corrections based on the FWER calculation can be implemented through single-step and step-down methods. In the former category all of the $P$ values are corrected by applying the same adjustment to all the biomarkers. The best known and simplest single-step correction procedure is the Bonferroni correction. In this method the corrected $P_i$ value, $correctP_i$, is equal to $P_i \times g$, and one rejects those hypotheses where $correctP_i < \alpha$.

Step-down methods by Holm (step-down Bonferroni) (Ewens and Grant, 2005) and Westfall, Young and Wright (1993) are less conservative than Bonferroni's corrections in the sense that more potential detectable differences may be detected, that is more hypotheses are rejected, by applying different adjustments to the different biomarkers. In the Holm correction method, the $P$ values are ranked from the smallest (top) to the largest (bottom), and they are corrected as follows. The top $P_j$ value is corrected by multiplying it by $g$ (i.e. $correctP_j = P_i \times g$), the second $P_j$ value is corrected by multiplying it by $(g - 1)$, the third $P_j$ value is corrected by multiplying it by $(g - 2)$, and so on, until no more hypotheses can be rejected.

The Westfall and Young method offers more statistical power than Bonferroni and Holm's method, but it implements a permutation procedure to estimate the distribution of $P$ values. As in the other methods, $P$ values are calculated and ranked for each biomarker using the observed data. A permutation method generates a 'pseudo-dataset' by randomly shuffling samples across the (control and case) groups in the original data. New *randP* values are estimated for all the biomarkers in this pseudo-dataset, and the minimum *randP* value is retained and compared to each of the observed $P$ values in the original dataset. This process is repeated thousands of times, and the corrected $P$ value, $correctP_j$, for gene $j$, is the proportion of pseudo-datasets where the minimum *randP* value was smaller than the observed $P$ value. More details about the mathematical implementation of these methods can be found in (Ewens and Grant, 2005) and (Glantz, 2001).

FDR methods for multiple-hypotheses testing aim to increase the power of statistical testing in comparison to FWER methods. But instead of simply allowing the possible occurrence of more false positives or of controlling the FWER, FDR methods estimate (or control) the number of potential false positive predictions that one would be prepared to accept. Thus, in a typical microarray data analysis involving thousands of genes, the researcher can have an estimate of how many genes would be false positive predictions, out of those found to be differentially expressed. For example, suppose that 20 000 genes were analyzed and that the statistical test reported 100 genes as differentially expressed, then a FDR = 50% indicates that 50 out of those 100 genes are expected to be false positive ('significant') predictions.

The original procedure for estimating FDR was proposed by Benjamini and Hochberg (1995). As in the FWER methods, the FDR correction method requires the estimation of a $P$ value for each gene, that is $g$ tests under $g$ independent, null hypotheses. These $P$ values are ranked from the smallest to the largest value, with ranking values $i = 1, \ldots, g$. The largest $P$ value is not corrected. Subsequent $P$ values are corrected by multiplying by $g$ and dividing them by their corresponding ranks. The $Ho$ is rejected for those corrected values that fall below the significance level, $\alpha$.

In order to address some of the limitations of this procedure, such as statistical assumptions about independence that are difficult to justify (Ewens and Grant, 2005), several techniques based on distribution-free or permutation methods have been proposed. One such permutation-based method for estimating the FDR is the Significance Analysis of Microarrays (SAM) proposed by Tusher *et al.* (2001), which have become a well-known analysis tool in the microarray research community. As input, this method accepts a data matrix, **D**, encoding the expression levels of $g$ genes across $s$ samples, belonging to two classes. The outputs are a list of differentially expressed genes and an estimation of the FDR. For each gene, a statistic value, $d$, similar to the $t$-statistic, is estimated (Tusher *et al.*, 2001; Ewens and Grant, 2005). Samples in **D** are permuted and a number, *numPer*, of random permuted datasets are obtained. Each dataset can be identified by an index, *per*: 1...*numPer*, with the original dataset representing the first dataset, $per = 1$. In each permutation, a $d$ statistic, $d_{per}(i)$, for each variable, $i = 1, 2, \ldots, g$, is calculated. For each permutation, the genes are ranked according to their corresponding $d$ values from the largest (top first) to the smallest (bottom). These ranked lists of $d_{per}(i)$ values represent the columns in a matrix of $d$ values derived from each permutation and variable. Note that a row, $i$, in this matrix will not necessarily correspond to gene $i$. Average $d$ values, $d_{avg}(i)$ from the entries in the $i$th row are calculated. This is followed by the calculation of the difference between the $d(i)$ values from the original data and $d_{avg}(i)$ for each gene. The SAM procedure also introduces the parameter, $\Delta$, to define when a gene, $i$, should be considered as potentially 'significant'. Thus, if the difference between a $d(i)$ and $d_{avg}(i)$ is greater than $\Delta$, one can state that the gene is differentially expressed. Larger $\Delta$ values will tend to reduce the number of false positives, while small $\Delta$ values will generate larger numbers of false positives.

The FDR is estimated based on a permutation procedure. First, for each permutation, SAM estimates the number of genes with a $d_{per}(i)$ value below the critical value $a$, or greater than critical value $b$. The former value refers to the largest negative value, $d_{per}(i)$, amongst those genes defined as (significantly) differentially expressed. The latter is the smallest positive value, $d_{per}(i)$, from the set of differentially expressed genes. The average of such numbers is calculated over all the permutations. This average value divided by the actual number of genes defined as 'significant' in the original data gives the approximation of the FDR. Additional mathematical details of the calculation of this and related versions of the procedure are available in (Ewens and Grant, 2005).

Recent research has discussed possible limitations and misinterpretations of comparisons of different studies based on FDR analysis. For example, it has been suggested that prediction inconsistencies across different gene expression data investigations, but using identical data, could be prevented by reporting additional statistical information, such as probability and expression ratios (Higdon, van Belle and Kolker, 2008). Moreover, it has been suggested that more consistent interpretations may be obtained by avoiding the use of pre-defined FDR thresholds. Other researchers (Jiao and Zhang, 2008) have argued that the standard permutation methods may overestimate the FDR, and have proposed variations of the test to address this problem. Comprehensive mathematical coverage of techniques for multiple-hypotheses testing can be found in the works of Dudoit and van der Laan (2008), which illustrate different applications using genomic data.

## 2.4 Correlation

The correlation between two variables (e.g. between-biomarkers, or biomarker-outcome) reflects the level of association between the variables. Such an association can be computed using parametric and non-parametric techniques. The former assumes that the variables can be jointly modelled with a normal distribution. The latter does not make this assumption, and is based on the idea of comparing the value ranks of the variables.

Correlation techniques will typically detect linear associations between the variables. They tell the researcher whether the variables vary in a similar fashion. If two variables are perfectly (positively) correlated, this means that when one shows larger (or smaller) values the second will also follow the same tendency. Correlation values range from $-1$ through 0 to 1, indicating perfect negative, no association and perfect positive correlation respectively.

The Pearson correlation coefficient, $r$, is the most commonly applied correlation method. The mathematical formula of this parametric method combines information about the degree of dispersion of the variables (around their means) independently and the co-variation between the variables. A non-parametric version is offered by the Spearman rank correlation coefficient. If parametric assumptions are satisfied, the Spearman coefficient may not be as powerful as Pearson's in detecting existing associations. Data transformations, such as log transformations of the variable values, may facilitate clearer visualization of correlation. Correlation values may be reported together with $P$ values, which are used to interpret the strength of the observed value, that is an estimate of the probability of observing such a correlation value by chance. Mathematical details and other examples using biomedical data are explained by Glantz (2001) and Crawford (2006).

Figure 2.1 illustrates correlations between two variables. In each graph, hypothetical values generated by two biomarkers (ordinate axis) observed across 10 samples (or time points) are compared. Lines linking data points were included to facilitate visualization. The top graph is an example of a positive correlation. In this case the Pearson correlation coefficient, $r$(biomarker 1, biomarker 2), is equal to 0.97. The bottom graph depicts a negative correlation, with Pearson correlation coefficient, $r$(biomarker 2, biomarker 3) = $-0.65$. In both examples stronger correlations are obtained when using the Spearman correlation (0.99 and $-0.80$ respectively). Figure 2.2 offers alternative displays of the association between these variables without incorporating information about sample number (or time dimension). Each scatterplot also includes a linear regression line fitted to the data, which estimates the linear association between two variables. The following section will further discuss regression analysis.

## 2.5 Regression and classification: basic concepts

Regression and classification applications are at the centre of most biomarker discovery investigations. In regression analysis one is interested in estimating quantitative associations between one dependant variable and one (or more) predictors. Dependent variables may represent a reference biomarker, an indicator of disease or any variable measuring a clinical response. Predictor variables or features may represent the potential novel biomarker(s) under investigation. There is a great variety of techniques for estimating the

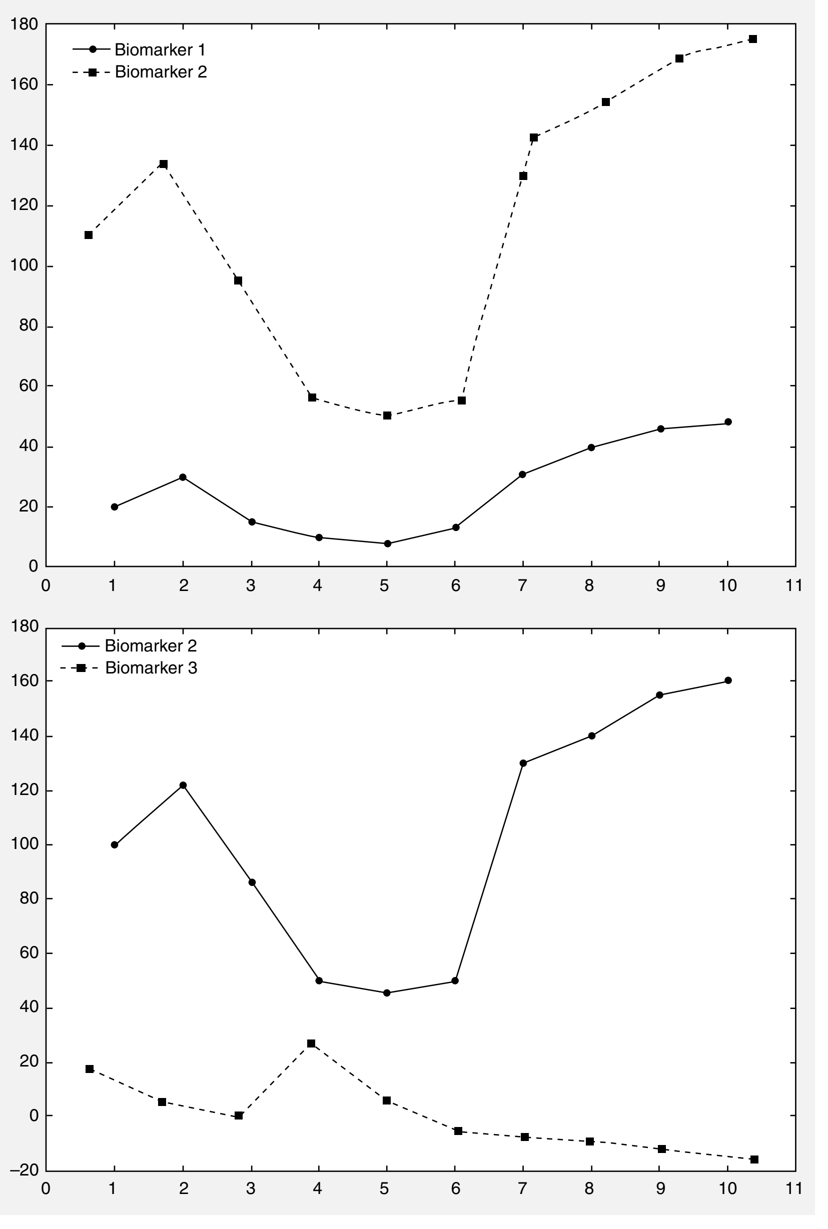


**Figure 2.1** Correlations between two variables. Top graph: is an example of a positive correlation, with Pearson correlation coefficient = 0.97 and Spearman correlation = 0.99. Bottom graph: negative correlation, with Pearson correlation coefficient = −0.65 and Spearman correlation = −0.80

optimal mathematical function (or model) to infer the values of a dependant variable using sets of features as inputs to the function. The next chapters will include examples of regression applications in biomarker discovery using different techniques from traditional statistical analysis and machine learning. However, a comprehensive coverage of

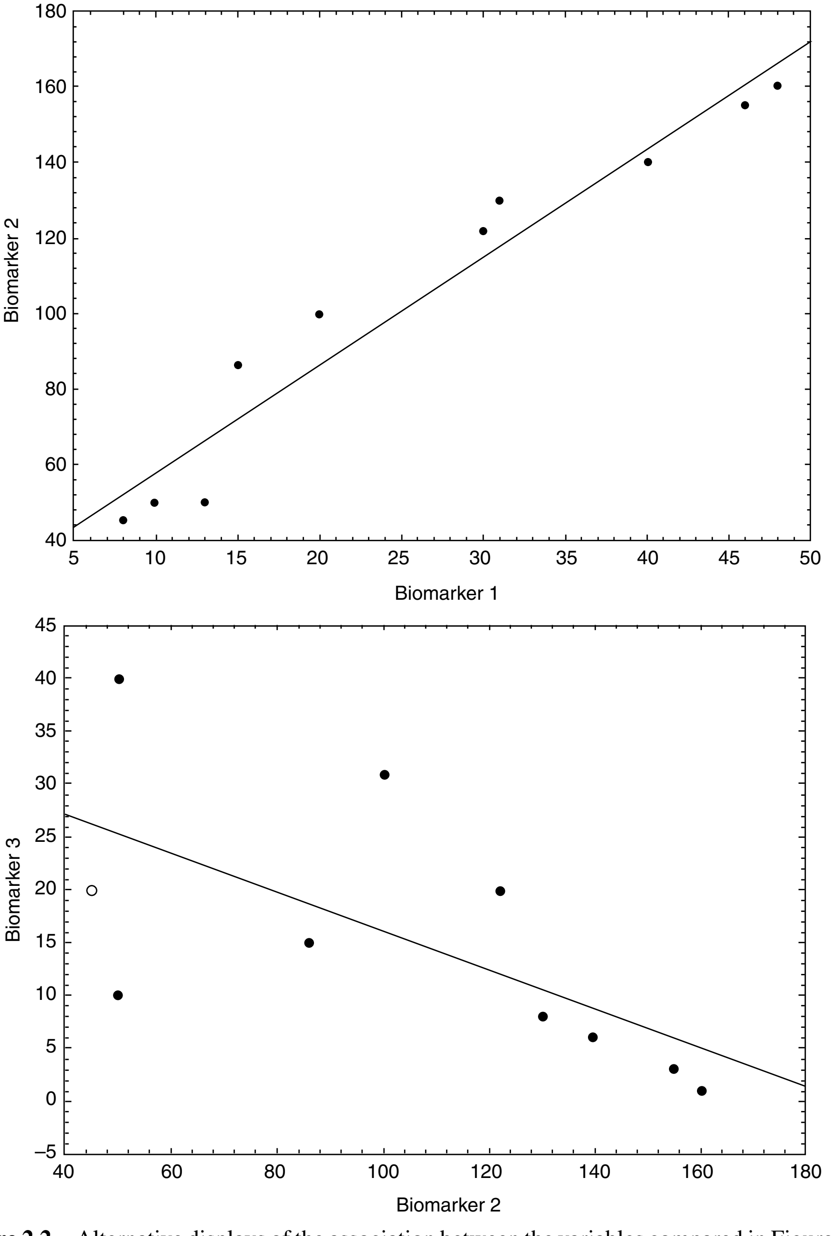


**Figure 2.2** Alternative displays of the association between the variables compared in Figure 2.1. Each scatterplot includes a linear regression line fitted to the data, which best estimates the linear association between two variables

regression techniques and their mathematical implementation are outside the scope of this book. Relevant papers and books have been published by Glantz and Slinker (2001) and Crawford (2006).

In the case of linear regression, the association between the dependant and predictor variables is estimated by fitting a linear function of the form: $y = \beta_0 + \beta_i x_i, \ldots, \beta_n x_n$, where $\beta_0$ is a constant value that represents the intercept of a (two-dimensional) regression function, $\beta_i$ are the regression coefficients (the function slope in a two-dimensional regression scenario), $y$ represents the Dependent variable, and $x_i$ represent the different predictors. This mathematical expression may be complemented by including residual errors that estimate the error incurred in predicting $y$ based on the resulting regression equation. Figure 2.2 shows examples of linear regression involving two variables. Each example graphically presents the linear function that can be used to predict the Dependent variable (ordinates) on the basis of a predictor (abscissa). In the top panel, for instance, the value of 'Biomarker 2' can be calculated as follows: (Biomarker 2) $= 29 + 2.9 \times$ (Biomarker 1). The coefficients can be estimated using different techniques, such as the least-squares method, which is the most used option in standard regression analysis (Glantz and Slinker, 2001; Crawford, 2006).

In classification applications one is interested in predicting a class, category or any research-meaningful label using sets of input features. Typical examples include the classification of patients into high and low risk on the basis of a number of potential biomarker values. In general there are two main categories of classification approaches: supervised and unsupervised classification. In the former category all the data need to be labelled prior to the analysis, that is the class labels need to be known in advance prior to the construction of a classification model. After the model has been built, 'unknown' or 'unlabelled' samples can be classified. In the area of unsupervised classification, information about sample classes is not analyzed or incorporated into the model construction process. Such information may be used to interpret the results and discover potentially relevant associations between groups of samples. Traditional data clustering falls into this category. In the next chapters the reader will find different applications in which clustering is used as part of exploratory data analysis or to support the search for potential biomarkers. Extensive literature on clustering algorithms and applications in bioinformatics has been published over the past 15 years in many journals, such as *Bioinformatics* and *BMC Bioinformatics*. Works by Teuvo Kohonen are recommended resources to obtain deeper insights into the problem of clustering for data visualization and classification (Kohonen, 2000). The next chapter will overview different data mining concepts, problems and approaches that emphasize the application of supervised classification.

## 2.6 Survival analysis methods

Survival data are obtained from studies in which the variable of interest is the time of occurrence of an event, such as death, a complication or recovery. In this case the time variable is referred to as the 'survival time' of an individual over a period of time. The occurrence of an event is also sometimes referred to as a 'failure'. This is an important source of information for many prognostic biomarker discovery investigations. These analyses estimate quantitative associations between potential biomarkers (and other features) and an event in a group of patients.

Censoring is a fundamental concept for understanding survival analysis. This term is used to define cases in which the survival times are not known exactly. This occurs when a patient does not experience a failure before the end of the investigation, or when the patient withdraws from it. An understanding of the meaning of the survival and 'hazard functions' is also essential to interpret survival data analysis.

The survivor (or survival) function, $S(t)$, describes the survival rates of the population sample under study. It is an approximation of the probability that a patient will survive (or experience an event) after a specific time, $t$. The hazard function, $h(t)$, indicates the potential failure rate at time $t$, given that a patient has survived until this time. In survival analysis publications both functions are typically reported with the survivor function represented by a 'step function' (Figure 2.3). At time $t = 0$, $S(0) = 1$, as no patient has experienced an event and the probability of surviving just after $t = 0$ is equal to one. For other times, $t$, the $S(t)$ values are estimated by calculating the proportion of patients surviving past time, $t$, that is a cumulative proportion of surviving patients at each time. The estimation of a survivor function can be carried out through the Kaplan-Meier (KM) method (Rao and Schoenfeld, 2007; Kleinbaum and Klein, 2005a). Figure 2.3 illustrates a hypothetical example in which the survivor functions of two patient groups, who undergo different treatments, are compared. A key method for comparing surviving curves, as well as for testing the hypothesis of no differences in terms of survival times, is the 'log-rank test'.

The hazard function, $h(t)$, can have different shapes according to the characteristics of the population investigated. For example, it will have a constant value, $\lambda$, for all times, $t$, in a group of healthy subjects. It will be a decreasing curve in the case of a group of individuals recovering from surgery, or an increasing curve when describing a patient group with a malignant disease. The most used hazard model is the Cox proportional

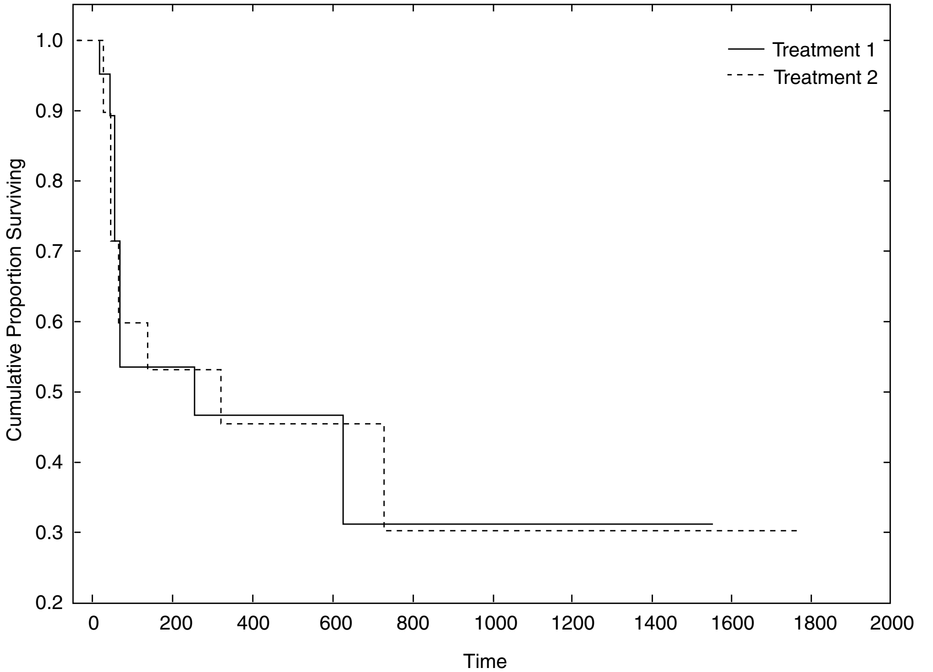


**Figure 2.3** Kaplan-Meier analysis: Hypothetical example in which the survivor functions of two patient groups, who undergo different treatments, are compared

hazards model, which is defined by the following function (Rao and Schoenfeld, 2007; Kleinbaum and Klein, 2005a):

$$h(t, x) = h_o(t)\exp(\beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k)$$

Where $h_o(t)$ is the baseline function, 'exp' is the exponential function, **x** is the vector of predictor variables, $x_i$. The parameters $\beta_{\mathrm{i}}$ are estimated by the maximum likelihood method from the data under analysis (Kleinbaum and Klein, 2005a).

Hazard ratios are used to make predictions about the hazard of an individual in relation to the hazard of a different individual. To illustrate its meaning, suppose that $x_1$ is the only variable under investigation. This variable may represent a risk factor, with binary values 0 and 1 encoding the absence and presence of the risk factor respectively. The hazard ratio for subjects with $x_1 = 1$ in relation to those patients with $x_1 = 0$ is equal to $\exp(\beta_1)$. In general, hazard ratios are estimates of relative risk, that is the risk ratio between an experimental (e.g. treatment group) and a control group. The Cox model is also used to estimate survival curves that adjust for the predictor variables included in the model. Like the KM curves, adjusted survival curves are step functions. Readers interested in a comprehensive description of survival analysis may refer to the works by Kleinbaum and Klein (2005a), which include practical examples using different statistical packages.

## 2.7 Assessing predictive quality

The quality of the predictions made by a biomarker-based model can be estimated by different quantitative indicators. The predictive ability of regression models can be summarized by error indicators, such as absolute errors, the mean squared error, or the root-mean-square error between the real and predicted values. In the case of classification a wider range of indicators are available that can be applied to different biomarker research applications. This section concentrates on prediction quality indicators for classification applications.

Each potential biomarker or prediction model can be assessed on the basis of its capacity to distinguish between experimental (case) and control subjects. This can be done by estimating their true-positive rate (TPR), that is the proportion of case samples that are classified as positive predictions, and the false-positive rate (FPR), that is the proportion of control samples that are incorrectly detected as positives. These and other indicators are derived from the basic error measures defined in Table 2.3. Based on such measures, one can define different global indicators of quality, as shown in Table 2.5.

Sensitivity and (1-Specificity) are synonyms for TPR and FPR respectively. When a biomarker produces continuous numerical scores (e.g. concentration values) different prediction thresholds (PT) can be defined to assign a sample to the positive class. That is, a sample is assigned to the positive class if, for example, the biomarker concentration is above a PT value. In this case a receiver operating characteristic (ROC) curve can be used to visualize the predictive ability of the biomarker (or combination of biomarkers integrated into a model) for different PT values (Swets, 1988). Each point on a ROC curve represents the TPR vs. FPR for a specific PT value. For example, if the output of a prediction model is a biomarker concentration value from 0 to 10, one could define PT

**Table 2.5** Definition of important indicators of classification performance

| Indicator | Definition | Meaning |
|---|---|---|
| True positives | TP | Number of positive cases correctly classified |
| True negatives | TN | Number of negative cases correctly classified |
| False positives | FP | Number of negative cases incorrectly classified |
| False negatives | FN | Number of positive cases incorrectly classified |
| Accuracy | Acc = (TP + TN)/ (TP + TN + FP + FN) | The proportion of all cases correctly classified |
| Sensitivity (Recall) | Sensitivity = TP/(TP + FN) | The proportion of true positive cases correctly classified |
| Specificity | Specificity = TN/(TN + FP) | The proportion of true negative cases correctly classified |
| Likelihood ratio (Positive class) | LR (+) = Sensitivity/ (1 − Specificity) | The likelihood of correctly predicting a positive case in relation to making the same prediction in a negative case (ruling-in disease) |
| Likelihood ratio (Negative class) | LR (−) = (1 − Sensitivity)/ Specificity | How likely a prediction model will label a truly positive case as a negative case in comparison with a truly negative case (ruling-out disease) |
| Precision | Precision = TP/(TP + FP) | The proportion of positive predictions that are actually positives. |

values at 0.5, 1, 1.5, . . ., 10. Thus, for each PT value the tested samples are classified, overall TPR and FPR are calculated and the resulting points are plotted.

ROC curves offer advantages over other statistical indicators, such as frequency-based scores: its scale-independence and the capacity to visualize FPR and TPR for different PT regions. The closer a ROC curve is to the upper and left axes of the plot, the more powerful the prediction capacity of the model under analysis. The closer a ROC curve is to the diagonal line connecting the upper-right and lower-left vertices of the plot, the poorer the performance of the model, that is the closer to a classification performance driven by chance.

Based on the (TPR, FPR) pairs obtained for different PT values, a ROC curve can be approximated by two main statistical techniques: Parametric and non-parametric (Shapiro, 1999). In the former case, one assumes that the data follows a specific statistical distribution (e.g. normal), which is then fitted to the observed test results to produce a smooth ROC curve. Non-parametric approaches involve the estimation of FPR and TPR using the observed data only. The resulting empirical ROC curve is not a smooth mathematical function, but a continuous series of horizontal and vertical steps.

| Sample No. | True class | Bio. conc. |
|---|---|---|
| 1 | Presence | 200 |
| 2 | Presence | 140 |
| 3 | Presence | 100 |
| 4 | Presence | 170 |
| 5 | Presence | 160 |
| 6 | Absence | 160 |
| 7 | Absence | 160 |
| 8 | Absence | 150 |
| 9 | Absence | 120 |
| 10 | Absence | 140 |

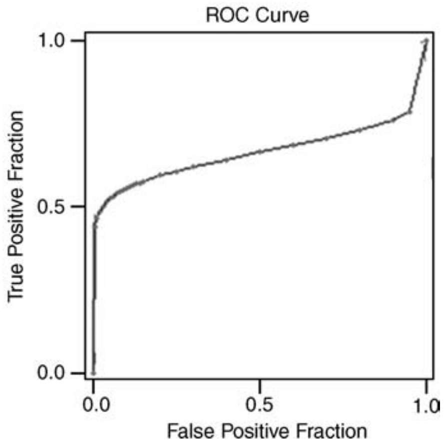


**Figure 2.4** Example of ROC curve obtained from testing data consisting of 10 samples, 2 classes: Presence and absence of a disease, and a prediction model based on the concentration values derived from hypothetical biomarker (Bio. conc.). Plots generated by the ROC calculator of Eng (2006) at Johns Hopkins University, using a curve-fitting parametric technique

Figure 2.4 illustrates a ROC curve obtained from predictions made on a testing dataset based on the numerical outputs generated by a hypothetical biomarker. If one evaluates a classification model (e.g. a classifier integrating multiple biomarkers) that generates numerical prediction scores for each class (e.g. probability values), one can use the scores produced for the positive class to define the PT values and estimate the different (TPR, FPR) value pairs (Figure 2.5). Figure 2.6 illustrates the comparison of (non-parametric) ROC curves derived from two classifiers independently tested on 400 samples in a hypothetical prediction analysis (200 samples/class). In this example, the classification model 'Csf-1' outperforms the model 'Csf-4' across all the prediction thresholds. The

| Sample No | True class | Prob. comp. |
|---|---|---|
| 1 | Complication | 0.9 |
| 2 | Complication | 0.4 |
| 3 | Complication | 0.5 |
| 4 | Complication | 0.7 |
| 5 | Complication | 0.6 |
| 6 | Recovery | 0.3 |
| 7 | Recovery | 0.6 |
| 8 | Recovery | 0.5 |
| 9 | Recovery | 0.2 |
| 10 | Recovery | 0.4 |

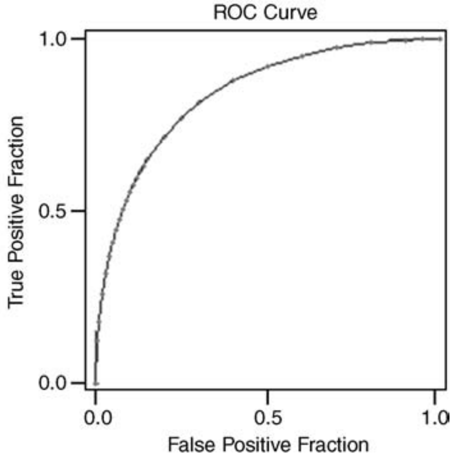


**Figure 2.5** Example of ROC curve obtained from testing data consisting of 10 samples, 2 classes: Medical complication and recovery, and a hypothetical classification model that assigns samples to classes according to numerical scores or probabilities. Plots generated by the ROC calculator of Eng (2006) at Johns Hopkins University, using a curve-fitting parametric technique

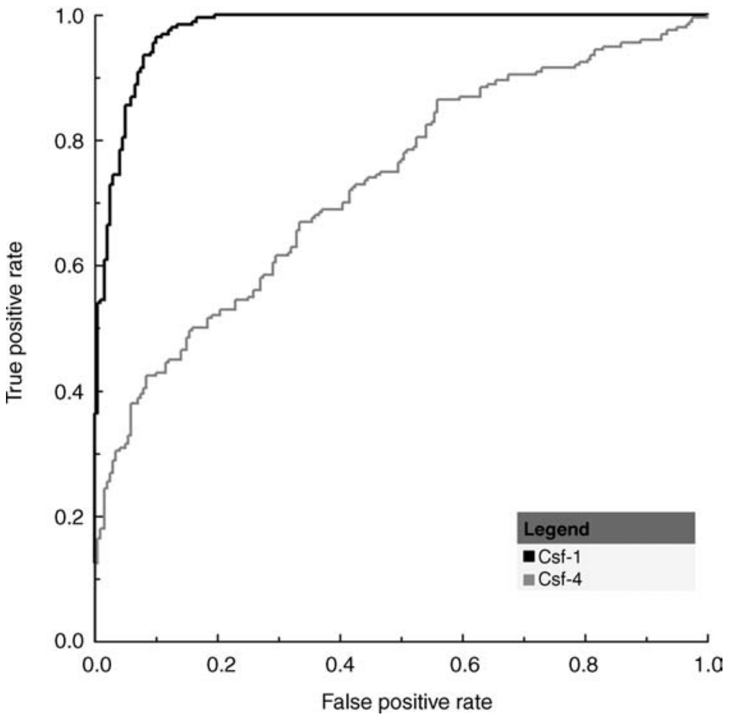


**Figure 2.6** Comparison of (non-parametric) ROC curves derived from 2 classifiers independently tested on 400 samples (200 samples/class). Plots were generated by the StAR system, at Pontificia Universidad Católica de Chile (Vergara *et al.*, 2008)

plots in Figures 2.4 and 2.5 were generated with the ROC calculator of Eng (2006), using a curve-fitting parametric technique. The ROC in Figure 2.6 was generated with the StAR system (Vergara *et al.*, 2008).

Once different biomarkers or prediction models have been evaluated individually, one should compare them on the basis of a quality indicator, for example a statistical test or the area under the ROC curve (Shapiro, 1999). This will allow one to rank the different models and establish priorities for subsequent analyses. If many potential biomarkers are being analyzed independently, it is also important to take into account the possibility of observing significant differences by chance only. Therefore, this type of analysis should include adequate adjustments or corrections, as well as estimations based on cross-validation (Chapter 3).

The area under the ROC curve (AUC), also known as the *c*-statistic, can be used to summarize predictive accuracy. To simplify, the AUC actually represents the probability that, given a pair of case and control samples randomly chosen, the case and control samples will be correctly classified as case and control samples respectively. It can also be defined as the probability of correctly ordering the risks of a pair of low and high risk samples (Pepe, Feng and Gu, 2007; Pepe, Janes and Gu, 2007). An AUC value of 0.5 is obtained when the ROC curve corresponds to the performance of a random classifier. The perfect accuracy is obtained when AUC = 1. As in the estimation of the ROC curves, there are different parametric and non-parametric ways to estimate the AUC (Shapiro, 1999). For example, the non-parametric estimate of an AUC can be implemented with the Wicoxon rank-sum test (Zou, O'Malley and Mauri, 2007).

Because the AUC is a global measure of predictive quality for a full range of PT values, and because ROC curves may intersect when comparing different models, the analysis of partial AUC values may facilitate a more objective interpretation. This will allow the researcher to concentrate on differences found in sections of the PT scale or for specific sample sub-groups (Zou, O'Malley and Mauri, 2007). Researchers have also demonstrated that precision-recall curves may provide a more accurate view of the predictive quality of a model when dealing with highly imbalanced or skewed datasets (Davis and Goadrich, 2006). An imbalance dataset is one in which the number of control samples is much larger than the number of case samples, or vice versa. Other researchers have also argued against the idea of over-emphasizing the importance of AUC values to compare prediction models. For instance, it has been shown that relatively strong biomarkers may have limited impact on changes in the AUC despite their capacity to significantly contribute to make better predictions (Cook, 2008). Pepe, Feng and Gu (2007) and Pepe, Janes and Gu (2007), have argued that AUC values are not relevant to a real clinical context due to the notion of pairing samples that it represents, and should not be used as the main evaluator of prediction quality in biomarker studies. In any case, it is important to evaluate models using different indicators that may reflect the impact of the models at the individual and population levels. Pepe, Feng and Gu (2007) and Pepe, Janes and Gu (2007) recommend an interpretation of AUC estimates in conjunction with a closer look at sensitivity and specificity results. Recent research also suggests that there is a need to include more graphical displays of predictive quality, including ROC curves, in primary diagnostic studies and systematic reviews (Whiting *et al.*, 2008). Different methods for estimating predictive quality have been extensively reviewed by Shapiro (1999), Cook (2008) and Pepe (2008).

In numerical risk assessment applications, calibration and reclassification evaluation methods are recommended. Calibration analysis focuses on the estimated probabilities or risk scores generated by a model. It compares the 'observed' and 'predicted' risk values in a group of patients (Cook, 2008). In a calibration model sub-groups of patients are defined to make these comparisons. The observed risk can be approximated by calculating the observed risk proportion within these sub-groups. For example, sub-groups can be defined by different user-defined intervals and deciles of estimated risk values (Cook, 2008). In a reclassification analysis the risk scores estimated by the compared models are grouped into different clinically-relevant categories. For example, four risk score categories can be defined as: below 5%, between 5% and 10%, between 10% and 20%, and larger than 20%. Given a reference model, RM, and a new prediction model, NM, a reclassification analysis compares the percentages of patients re-classified or assigned to different categories by NM in comparison to RM. The percentage of patients reclassified by NM provides the basis for estimating the potential prediction impact of this model (Cook, 2008).

## 2.8 Data sample size estimation

Biomarker discovery studies require careful planning and design. Samples are selected from relevant groups, standard experimental protocols are followed, and different randomization or data sampling procedures may be required. An estimation of the number of samples aims to ensure that enough data are acquired to allow the detection of statistical differences when such effects are indeed present in the population. Undersized and oversized studies can have major implications, not only in terms of economic costs. Costs involving unnecessary or potentially dangerous exposures to treatments and ethical conflicts may also arise.

Sample size estimation can be done with different commercial and free software tools. In the latter category, the programs developed by Lenth (2000) and Dupont and Plummer (2008) are recommended, which cover different applications and study designs.

The estimation of the optimal sample size of a biomarker study is a complex problem. Traditional sample size estimation techniques based on statistical significance tests may be useful references, but may be shown to be inaccurate in some biomedical studies. Other authors have suggested the use of computer-based simulations involving synthetic data, which recreate some of the characteristics of the real classification problem under consideration. Independently of the sample size estimation approach selected, the aim is to determine the number of samples required to meet basic assumptions about the data, hypothesis testing procedure to be applied and study outcomes. Sample size estimation should be seen as a context- and study-dependent problem.

One such approach to sample size estimation is based on the analysis of the statistical power of a hypothesis testing procedure. To do this, the user is typically required: (a) to specify *H0* and *Ha*, (b) to specify a hypothesis testing procedure, (c) to define the desired significance level, (d) to specify the effect size required to reject the hypothesis (e.g. expected difference between means), and (e) to estimate input parameters needed (e.g. group variances) to analyze the statistical power and determine the sample size meeting these specifications. Hence, the estimated sample size is the minimum number of samples required in each group to satisfy these assumptions and conditions (Lenth, 2001).

There is no standard or application-independent methodology to determine the above input parameters (Eng, 2003). In the case of the desired power, it is common to define a power of 0.80, with higher values increasing the estimated sample sizes. The definition of these parameters requires inputs from different researchers in a team. Meaningful estimates may also be obtained by looking at previous published research or by implementing pilot studies. In some teams, some researchers may be tempted to let the bioinformatician or biostatistician do the 'number crunching' by themselves. Nevertheless, the estimation of input parameters and targets should be carefully discussed by all the researchers with key responsibilities in the project. Bioinformaticians and biostatisticians can facilitate the determination of such values by posing questions to the wet lab scientists and managers in terms of expected scientific findings, costs and losses (economic and those incurred due to lack or excess of power), and upper bounds for sample sizes. Lenth (2001) offers practical advice on how to support study planning and communication for different scenarios, including those severely constrained by low budgets and management decisions. Sample size estimation requires close cooperation between statisticians and domain experts, and may involve asking many questions about the goals of the study, sources of variation, and the need for actually estimating sample sizes.

Different authors have presented formulae for estimating sample sizes in different types of experimental situations, goals, hypothesis testing techniques and requirements

(Eng, 2003). For example, diverse formulae have been recommended for different experimental microarray designs (Dobbin and Simon, 2005). Many of these approaches concentrate on the analysis of power involving individual genes. Such an approach is also, in principle, a valid starting point in studies involving multiple biomarkers and classification problems (Dobbin and Simon, 2005). However, in the case of classification models involving multiple biomarkers as inputs features, the 'optimal' sample size is also related to the optimal number of inputs (Hua *et al.*, 2005). And these factors will in turn depend on the classification model and data distribution under study.

A meaningful approach to estimating sample sizes in microarray studies involving multiple-hypotheses testing is the interpretation of previous findings derived from similar designs. Page *et al.* (2006) developed the PowerAtlas to allow researchers to estimate statistical power and sample sizes based on the interpretation of previous studies showing similar experimental characteristics. The PowerAtlas contains hundreds of experiments from the Gene Expression Omnibus (GEO), and allows users to specify the characteristics of their studies, and to select published datasets that satisfy the selection criteria. The analysis of previous studies and determination of sample sizes are based on the concepts of Expected Discovery Rate (EDR) and the proportion of true positives (PTP) (Gadbury *et al.*, 2004). These concepts are used by the PowerAtlas to guide the user in the selection of appropriate sample sizes. The EDR is the average power for all genes expected to be differentially expressed in the study. The PTP refers to the proportion of genes detected as differentially expressed amongst the set of genes truly differentially expressed in the data.

## 2.9 Common pitfalls and misinterpretations

Table 2.6 presents examples of common pitfalls and misinterpretations of basic statistical concepts in biomedical research, in general, and in the biomarker discovery literature, in particular. Guidance and key resources for further learning are provided for each aspect. These problems range from misunderstandings of fundamental definitions (e.g. statistical 'significance' and the meaning of $P$ values), misinterpretations of the purpose of statistical analysis (e.g. the meaning of sample size estimation outcomes), through the inadequate interpretation of different prediction quality indicators (e.g. accuracy interpreted as precision, and vice versa), to the lack of good practices in reporting prediction results.

One of the greatest challenges is to improve the understanding of the meaning of hypothesis testing results. For example, many researchers make excessive use of the term 'significant' without properly considering both statistical and research domain contexts. In this case, any findings that generate $P$ values below 0.05 are automatically defined and reported as 'significant'. Researchers may often be tempted to ignore those findings falling above this threshold, because they are simply 'not significant'. The key is to understand the meaning of the parameters and outcomes of a hypothesis testing procedure and to interpret them in the context of the goals, limitations and characteristics of the investigation. Another common mistake is to believe that a $P$ value represents the probability of 'not significance', or the probability that the null hypothesis is true. Researchers should at least remember that $P$ values actually measure the strength of the evidence found against the null hypothesis. Moreover, hypothesis testing results with

**Table 2.6** Examples of common pitfalls and misinterpretations of basic statistical concepts

| Aspect | Problem | Key resource |
|---|---|---|
| Hypothesis testing | What significance really means? Misinterpretation of *P* values | Sterne and Davey Smith (2001); Glantz (2001) |
| Hypothesis testing | Approximations of *P* value ranges are reported only. Sufficient information on statistic values and techniques applied is not presented | Glantz (2001) |
| Hypothesis testing | *P* values are not adjusted when testing multiple hypotheses | Dudoit and van der Laan (2008) |
| Sample-size estimation | Incorrect interpretation of estimates, lack of understanding of the need for size estimation in study planning | Davis and Mukamal (2006) |
| Correlation | Linking correlation to causation, misinterpretation of *P* value | Glantz (2001) |
| Risk indicators | Linking odds ratios to relative risks | Davies, Crombie and Tavakoli (1998) |
| Classification quality indicators | Inadequate interpretation of accuracy, precision, sensitivity, and specificity | Table 2.4 |
| Classification quality indicators | Misuse of AUC values | Shapiro (1999), Cook (2008), Pepe (2008) |

*P* values falling below the level of 0.05 may not always represent strong evidence against the null hypothesis.

Another possible incorrect conceptual interpretation is to interpret a sample size estimation procedure as the prediction of the number of samples needed to 'obtain significance' or to ensure 'significant results'. Also it is wrong to assume that statistical analyses are not sensitive to assumptions about the data and the study design under consideration. In addition, it is important to ensure that in all discussions there is a clear separation between the concepts of correlation or causation. Researchers may be tempted to conclude that causal relationships can be discovered solely on the basis of relatively high correlation values, especially when such evidence appears to support their expectations, hypothesis or previous research. Correlation and causal associations are different concepts. Although the former may be seen as a necessary condition to observe the latter, they can never be seen as synonymous definitions even in cases of very strong correlation.

Another concern is the possibility to misinterpret the meaning of *P* values associated with correlation estimations. Researchers should not interpret a correlation as very strong or 'significant' simply because the $P$ value associated with the correlation falls below a significance value of 0.05. Such $P$ values actually assess the possibility that the observed correlation value can be obtained by chance alone or when there is no actual correlation in the population, independently of whether the researcher considers it as a large or small correlation value.

When interpreting AUC values, researchers should bear in mind that this indicator is not a synonym of sensitivity or of the overall rate of samples correctly classified by a model. Moreover, there is a need to avoid an over-emphasis on AUC values as the most relevant indicator of predictive quality. The reporting of different quality indicators and judicious context-based interpretations should be given careful consideration.

Researchers should also consider the potential implications or sources of misunderstanding when interpreting indicators of the size of an effect, such as odds ratios (Glantz, 2001), which aim to estimate the effects of exposures or treatments in control-case studies. For example, it has been demonstrated that odds ratios may overestimate an effect size when odds ratios are interpreted as relative risks (Davies, Crombie and Tavakoli, 1998). The use of odds ratios to approximate relative risks may be misleading in studies with large effects on patient groups with high initial risks (Davies, Crombie and Tavakoli, 1998). On the other hand, it has been shown that odds ratios would not underestimate the relative risk in studies showing reductions in risk, that is odds ratios smaller than one (Davies, Crombie and Tavakoli, 1998).

# 3 Biomarker-based prediction models: design and interpretation principles

This chapter will introduce key techniques and applications for patient classification and disease prediction based on multivariate data analysis and machine learning techniques, such as instance-based learning, kernel methods, random optimization, and graphical and statistical learning models. An analysis of prediction evaluation, model reporting and critical design issues will be provided. This chapter will also discuss feature selection for biomarker discovery.

## 3.1 Biomarker discovery and prediction model development

Disease classification and risk prediction models are typically based on multivariate statistical models involving different predictive factors. These models can be implemented with mathematical functions, non-parametric techniques, heuristic classification procedures and probabilistic prediction approaches. However, multi-biomarker prediction models may not always be strongly correlated with a disease or phenotype, or may not fully reflect inter-individual variability associated with the prediction output.

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

Moreover, the directed incorporation of biomarkers from relatively well-studied functional pathways may introduce bias and may not account for the functional interdependence and diversity inherent in complex diseases.

Typical examples of clinical classification systems based on biomarkers are: classification of healthy vs. diseased patients, the classification of survival/death outcomes, and the prediction of poor/good prognosis after therapeutic intervention. In a typical prediction model design problem, $m$ independent samples or observations are available, which are described by $n$ random variables. Each sample can be encoded by a feature vector $\boldsymbol{x_i}$ consisting of $n$ feature values, $x_{i,j}$, $i = 1, \ldots, m$; $j = 1, \ldots, n$. The features may represent gene expression values, clinical risk factors or the intensity values from proteomic or metabolomic data. Additionally, each sample, $\boldsymbol{x_i}$, is associated with an outcome value $y_i$, which may encode a continuous or discrete value, such as disease status or survival time. As defined in Chapter 2, the type of outcome to be predicted defines the type of prediction task or model to be selected: either regression or classification models. A summary of relevant statistical and machine learning methodologies that can be applied to construct regression and classification models is presented below, together with discussions on their strengths and limitations for biomarker discovery.

A prediction model offers a rule (e.g. mathematical function, algorithm or procedure) that uses information from $\boldsymbol{x}$ to predict $y$ in each observed sample. But more crucial, the goal is to use this model to predict the unknown outcome, $y_k$, for any observed $\boldsymbol{x}_k$ sample. Predictive generalization is the capacity to make accurate predictions of unknown outcomes, $y_k$, for different (testing) samples, $\boldsymbol{x}_k$, outside the set of samples, $\boldsymbol{x}_i$, used for building the prediction model. When this capacity is not achieved, the model is said to have over-fitted the learning or training dataset. Thus, the goal is to build a model that maps the input to the outcome information space with maximum generalization potential.

Independently of the prediction task, model development typically involves two major phases: Model learning (sometimes also referred to as training) and model evaluation (or testing). The learning phase allows the construction of the prediction models using a learning data set. This phase allows the 'learning' of the data characteristics and concept to be classified by fitting a mathematical representation or model to the training dataset. The predictive performance derived from the training phase is not always a reliable indicator for model evaluation purposes. Moreover, an 'over-fitting' of the model to the training data will likely result in a lack of predictive generalization in subsequent evaluations.

## 3.2 Evaluation of biomarker-based prediction models

Cross-validation (CV) and independent validation are the major approaches to estimate model prediction performance. These approaches may use different quantitative indicators of predictive quality as introduced in Chapter 2, such as $P$ values, AUC values, sensitivity, specificity, and so on.

CV comprises the selection of disjoint and randomly selected training and testing datasets, which are used for model training and testing independently. This data sampling and model construction process can be repeated several times using independent training-testing partitions. Overall performance is estimated by aggregating the performance originating from the different training-testing iterations. Despite the validity and robustness of this methodology, overestimation of predictive performance may be obtained when dealing with small datasets. Sub-optimal CV may be even more critical in prediction models that combine feature selection and classification. When reporting evaluation results involving multiple models, which can be based on different techniques or input features, it is crucial to report all the relevant evaluations implemented.

Independent validation consists of evaluating a prediction model on a completely different dataset, which should be independently generated and derived from a different set of biological samples. Limitations, constraints and key design factors should be carefully considered before its implementation. For example, it is essential that different teams participate in the data generation and analysis independently. Validation samples should also be independent from those used in the model construction (and CV). Moreover, as in the case of CV, the reporting of independent evaluations should be detailed and sufficient to facilitate unbiased interpretation and reproducibility. CV and independent validation may be seen as complementary methodologies in biomarker discovery research, with the former preceding the latter.

The best known data sampling techniques for estimating model predictive performance are: the traditional hold-out method, $k$-fold CV, leave-one-out CV and bootstrapping.

In the hold-out method a single learning-testing partition is defined using a predetermined proportion of samples in each set, for example two-thirds of the data are used for building the model (learning) and one-third is used for testing the model.

The $k$-fold CV method randomly assigns $m$ samples (i.e. observations, patients) to one of $k$ data partitions of equal size. In each learning-test iteration, the model is built using $(k-1)$ partitions, and is tested on the remaining partition. The overall prediction performance indicator (e.g. based on accuracy or AUC values) may be estimated as the average of the performance values obtained from each test fold.

The leave-one-out CV (LOOCV) is a version of the $k$-fold CV. Given $m$ samples (cases or patients), a classifier is trained on $(m-1)$ samples and tested on the case that was left out. This learning-testing process is repeated $m$ times, until every sample in the dataset has been included once as a testing sample. The model predictive performance is estimated using the prediction results from the $m$ test predictions.

The traditional bootstrap method is based on the idea of generating a training dataset by sampling with replacement $m$ times from the available $m$ cases (Efron and Tibshirani, 1993). The classifier is trained on this bootstrap dataset and then tested on the original dataset. This process is repeated several times, and the estimation of the prediction model's performance is based on the average of these test estimates. Several versions of the bootstrap have been proposed (Good, 2006), such as the leave-one-out bootstrap, the 0.632 bootstrap and the 0.632 +, which display different levels of bias and robustness in different application domains (Efron and Tibshirani, 1997).

The selection of a data sampling method for prediction model performance assessment is a context- and application-dependent problem. However, there is agreement that the hold-out method should be avoided. This is because of the highly biased and inexact estimations that this method produces, independently of the prediction model or algorithm applied. Recent empirical research has confirmed that the LOOCV tends to generate estimations of prediction performance with small bias, but with elevated variance. It has been suggested that the LOOCV and 10-fold CV may offer both the smallest bias and lowest mean square error in classification problems using different approaches (Molinaro, Simon and Pfeiffer, 2005). But also it should be noted that in general the more data available, the less significant the differences amongst these methods in terms of their capacity to estimate the 'true' prediction performance of a model.

## 3.3 Overview of data mining and key biomarker-based classification techniques

There is no universally accepted definition of the term 'data mining'. However, there is agreement that it refers to the application of different computational and statistical techniques to support knowledge discovery based on a better understanding of the data. Thus, the main goal of data mining is to make data meaningful. In the biomarker discovery context, this means the identification of more powerful and biomedical-meaningful biomarkers. Data mining also offers tools for the interpretation and evaluation of the resulting models and predictions, as well as the methods for supporting the implementation of prediction explanation mechanisms.

Different statistical and computational learning approaches can be used in biomarker data mining (Hastie, Tibshirani and Friedman, 2001). The difference between data mining and traditional statistical data analysis is that the former aims to identify unknown patterns, relations and meanings that could not be obtained by applying traditional statistical methodologies alone. Therefore, data mining is seen as an area that combines different approaches originating from different fields, ranging from statistical data analysis, computational intelligence and information visualization, amongst others.

Different problems require the application of different approaches and specific prediction models. Key selection factors that need to be considered could be the capacity of the method to deal with incomplete or missing data, or with different types of data (e.g. categorical or numerical data only or both), and the computational costs involved in implementing and deploying the models (Bellazzi and Zupan, 2008). In biomarker-based decision making support, it is also important to consider the potential of a method to allow the user to interpret prediction outcomes at different levels: visualization, natural language, graphical or probabilistic, and so on. Moreover, it is evident that any biomarker discovery process should allow the development of prediction models with a generalization capability, that is the ability to make 'correct' predictions when tested on unseen or unlabelled samples.

A brief introduction to some of the best known techniques for biomarker discovery and classification is presented here. This includes some of the most used data mining techniques for classification and regression tasks in different application domains, according to a 2006 poll that involved the opinions from data miners from academia and industry (poll results available at www.kdnuggets.com). Deeper descriptions of their implementation, as well as their applications in biomarker discovery studies in different biomedical areas and using different types of 'omic' data, will be illustrated in the next chapters. Chapter 9 discusses bioinformatic infrastructures, including software packages, for biomarker data mining.

### *Decision trees*

Decision trees allow hierarchical representations of the relationships between features in a dataset, such as potential diagnostic biomarkers (Breiman *et al.*, 1984; Quinlan, 1993). Such relationships and the values of each feature allow the model to classify new samples or cases. Before discussing the construction of decision trees, let's first overview how these structures are used to make specific predictions on a hypothetical testing dataset. Figure 3.1 illustrates a decision tree obtained from a training dataset consisting of four biomarkers: gene X, protein Y, clinical risk factor Z and protein Y2; and two hypothetical diagnostic classes: C1 and C2. The 'nodes' represent predictive features and specific feature values. The 'leaves' contain the samples that are classified under a specific category, which satisfy the different feature value conditions that are obtained by traversing the tree from the root node to the leave. In this example the nodes are represented by solid-line rectangles, and the leaves by dashed-line rectangles. Hence, each tree node represents a question about features values, whose answer will in turn allow the selection of the next node in the tree, from the top to the bottom.

This collection of questions-answers allows one to transverse the tree through a path that starts at the 'root node' and ends at one of the nodes without children nodes, that is the tree leaves. Typical questions include inequality questions, such as 'is the value of feature $y$ greater than $c$?', or questions involving more complex logical or mathematical combinations of features. In Figure 3.1, each leaf indicates the percentage of samples that fall into the leaf and that belong to the different classes under investigation. This information allows one to make predictions on new samples based on a majority vote or a probability distribution over the classes predicted. Thus, a sample is assigned to the class associated with the leaf reached. Note that different graphical representations can be used in different publications and software tools.

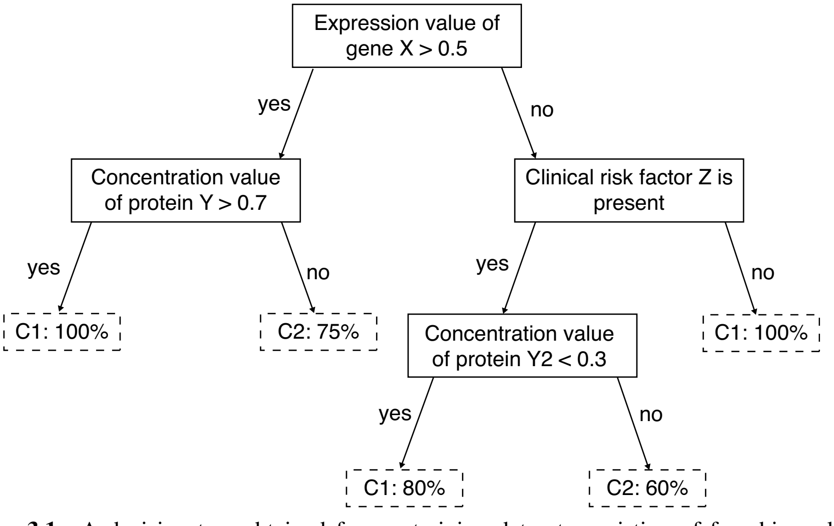


**Figure 3.1** A decision tree obtained from a training dataset consisting of four biomarkers: gene X, protein Y, clinical risk factor Z and protein Y2; and two hypothetical diagnostic classes: C1 and C2

Suppose that two new samples, $\mathbf{s}_1$ and $\mathbf{s}_2$, are represented by a vector encoding the expression value of gene X, the concentration value of protein Y, the clinical risk factor Z, and the concentration value of metabolite Y2. And suppose that these samples have the following feature values:

$$\mathbf{s}_1 : (0.6, 0.8, \text{absent}, 0.2)$$

$$\mathbf{s}_2 : (0.2,\ 0.1, \text{present}, 0.4)$$

The first case will be assigned to the class C1 because in this sample the expression value of X is greater than 0.5 and the concentration value of Y is greater than 0.7. This means that $\mathbf{s}_1$ falls into the left-most leaf, which groups a majority of samples that belong to C1 (100% of the samples in this leaf). Sample $\mathbf{s}_2$ is assigned to class C2 because in this sample the expression value of X is not greater than 0.5, the clinical risk factor is present and the concentration value of Y2 is not lower than 0.3. Thus, this sample falls into the leaf in which the majority (60%) of its samples belong to class C2.

The construction of a decision tree from a training dataset is based on the principle of data partition with regard to each feature and the satisfaction of different classification criteria. The main idea is to generate a node every time specific values, that is ranges or intervals, from a given feature maximize a classification criterion. A recursive partition process, starting with the selection of a root node, ensures that all samples in the training dataset will be assigned to one of the tree leaves. A question is assigned to each node (feature) if it allows a relative good split of samples in the training dataset. An optimum partition is selected based on measurements of ‘heterogeneity’ or ‘impurity’ of the class distribution associated with each putative new node. Two of the best known measures of impurity are entropy and the Gini index. Several decision tree construction algorithms are available, which differ in the way they infer and refine the tree structure or in how they evaluate prediction quality. Different versions based on the C4.5 algorithm, such as See5 and the CART algorithm (Breiman *et al.*, 1984; Quinlan, 1993), are available in public and commercial data mining packages (Chapter 9).

As in the case of other data-driven prediction models, an important challenge is to prevent a decision tree from over-fitting the training data. In decision trees this may be prevented by limiting the size or complexity of the resulting tree. A typical approach to ‘tree pruning’ consists of stopping tree growth, that is node splitting, when additional questions do not contribute to an increase of node purity above a pre-determined threshold value. Decision trees can also be adapted to regression applications. Recent research have also shown the advanced predictive capability of ensembles of decision trees, known as random forests, in different molecular profile classification applications. Alternative explanations and more detailed introductions to tree construction algorithms authored by Witten and Frank (2005) and by Kingsford and Salzberg (2008) are recommended.

### *Random forests*

Random forests represent an approach to classification based on the randomized construction and predictive integration of multiple decision trees (Breiman, 2001). Such a diversity of decision trees is obtained by training different trees on modified versions of the original training dataset. These training datasets can be generated by sampling with replacement from the original training data. Diverse trees can also be generated by selecting only a small, random subset of features during the construction of each tree. This diversity of trees and predictions for each sample is the key to contribute to a reduction of the likelihood of making incorrect classifications. Random forests exploits the assumption that the greater the prediction agreement between 'experts' (i.e. diverse, well-trained decision trees) with different views of the same problem, the greater the chances of making good predictions.

### *Logistic regression*

Logistic regression is a member of the family of generalized linear models (Kleinbaum and Klein, 2005b). Logistic regression is commonly used in classification problems involving binary responses or outcomes, such as the presence or non-occurrence of a disease. In this technique a 'logistic function' is fitted to the data (Kleinbaum and Klein, 2005a, 2005b). The resulting mathematical function is then used to estimate the risk of clinical outcomes or to predict prognostic categories of new samples based on sets of biomarkers. The fitted logistic function includes different regression coefficients associated with each biomarker. These coefficients also offer a quantitative estimation of the strength of the independent association between a biomarker, such as a risk factor, and the prediction outcome under investigation.

### *Artificial neural networks*

The application of artificial neural networks for biomedical decision support has been extensively investigated over the past 30 years. It includes a great variety of prediction techniques inspired in basic concepts and mechanisms observed in natural neural networks (Hastie, Tibshirani and Friedman, 2001; Russell and Norvig, 2002). In general, the construction of these models requires the definition of 'network architectures' and 'learning parameters'. The former refers to the structure of the networks: number of processing units (neurones), connections between them, number of layers of processing units, number of inputs and outputs, and the types of mathematical functions used by the network to process inputs and signals in the network. The user is required to select learning parameters that are used to initiate and guide the training of the network, such as the number of training epochs.

The network training can be implemented by using different standard algorithms according to the network type and prediction goals. One example of such algorithms is the 'back-propagation' algorithm, which is applied to train 'feed-forward multilayer perceptron' networks (Hastie, Tibshirani and Friedman, 2001). These algorithms use the training data to find a set of optimal parameters, for example weights associated with each input or signal processed by individual neurones, which best fit the input data (e.g. biomarkers) to the space of class values. Neural network training involves multiple steps of input presentation, output evaluation and model parameter adjustments. As in the case of other data classification techniques, the resulting prediction model is ready to be tested on an independent, testing dataset, and its predictive performance can be estimated by using CV.

### *Support vector machines*

Support vector machines currently represent some of the most powerful classification models. A great variety of applications to the classification of gene expression profiles and other types of biomedical data have been reported in high-impact factor journals in bioinformatics, biotechnology and biomedical sciences. Their popularity has been consolidated, in part, by their proven capacity to learn accurate prediction models and to deal with the 'curse of dimensionality' in different applications (Mjolsness and DeCoste, 2001; Hastie, Tibshirani and Friedman, 2001). These techniques are based on the principle of finding an optimal, linear separation of samples belonging to different classes, typically two classes, through the definition of a 'hyperplane' (Boser, Guyon and Vapnik, 1992). An optimal hyperplane is defined as the hyperplane that maximizes the separation between samples from different classes and its own separation to the nearest class-specific group of samples. The set of samples that is closest to the optimal hyperplane are known as the 'support vectors'. Thus, the support vectors define maximum margin hyperplanes. In practice, linear separation is not always possible and classification errors cannot be prevented. To address this requirement, a support vector machine defines a 'soft margin', which allows some samples to be misclassified, that is the samples are allowed to fall into the 'wrong side' of the separating hyperplane. The soft margin is a parameter that can be controlled by the user. This selection aims to estimate a margin size flexible enough to keep margin violations (classification errors) to the minimum.

Another problem is that in some datasets not even the introduction of a soft margin would allow an optimal linear separation due to the presence of many overlapping areas or mixtures of samples from different classes. Support vector machines aim to overcome this obstacle by transforming the original input data (e.g. set of biomarkers) into a higher-dimensional feature space. This transformation is implemented by applying non-linear mathematical functions on the original data. These mathematical functions are known as the 'kernels'. Examples of non-linear kernels are polynomial with degree greater than 1 and radial functions. In some applications this projection into a higher-dimensional space facilitates the identification of hyperplanes that may be used to approximate a linear separation between the samples.

For example, a dataset that could not be linearly separated on a one-dimensional space may be projected into a two-dimensial space to achieve the desired discrimination. Nevertheless, the higher the dimensionality of the resulting transformation, the higher the number of potential solutions that can be found to separate the data, that is the classification problem becomes harder to solve due to the curse of dimensionality (Noble, W.S., 2006a and Noble, D., 2006b). Thus, if a dataset is transformed by using a very high-dimensional kernel function then it is likely that any hyperplane solution will over-fit the data, that is the separation boundaries between samples will be very specific. The selection of the optimal kernel may be a complex and time-consuming task. Software packages allow the user to choose a kernel from a list of well-known options. The optimal kernel may be selected by simple trial and error, or guided by the results obtained from CV procedures.

Figure 3.2 illustrates the concepts of hyperplane, support vectors and linear separation using a hypothetical example of diagnostic classification. Figure 3.3 (top) shows a dataset consisting of samples described by a single feature, that is one-dimensional space, which could not be linearly separated by a single hyperplane. A typical example

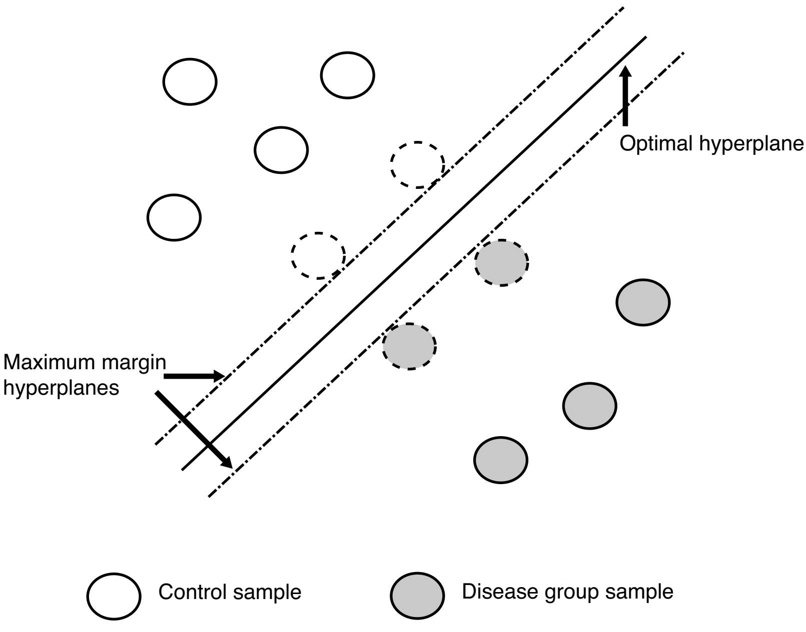


**Figure 3.2** Visualization of a hypothetical diagnostic classification of samples using the support vector machine technique. Samples are linearly separated with an optimal hyperplane. The support vectors are indicated with dashed circles

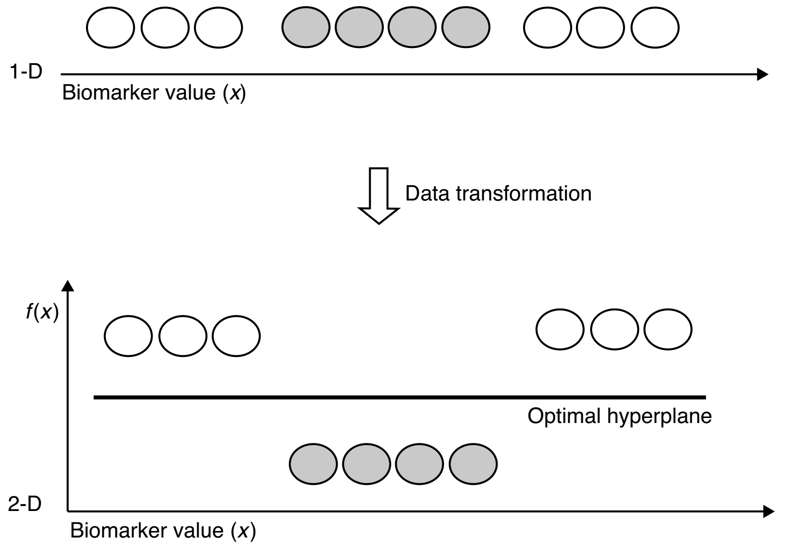


**Figure 3.3** Linear separation of samples after transformation of the original dataset (1-dimensional space, with $x$ representing the values of a hypothetical biomarker) into a 2-dimensional space, using a polynomial function $f(x)$

could be a dataset in which each patient is described by a single biomarker value, which in its original form does not provide sufficient information to detect non-linear relationships. This figure also explains how the transformation of this dataset into a higher-dimensional space (two-dimensional), using a simple polynomial function, would aid in the linear separation of the data and a more accurate classification.

### *Naïve Bayesian classifier*

Based on the assumption of statistical independence between features given the class to be predicted, the naïve Bayesian classifier estimates the probability that a sample belongs to the classes under consideration (Jensen and Nielsen, 2007). These predictions are based on the calculations of different conditional probabilities relating features values and classes in the (training) data, for example the probability that a sample belongs to 'class A' given that the 'feature $x$' has a value equal to 0. In the test mode, and under the conditional independence assumption, this technique allows the calculation of the probability values that a given sample given its feature values belong to each of the classes investigated. These estimations are based on the multiplication of the corresponding conditional probabilities calculated using the training set, which are stored in probability tables assigned to each feature. The sample is assigned to the class that reports the highest conditional probability. This technique has been widely applied to different domains and has been used as a benchmark classifier in different applications (Witten and Frank, 2005). A naïve Bayesian classifier only takes into account conditional dependencies between features and the class to be predicted. This makes the naïve Bayesian classifier the simplest type of Bayesian network model.

### *Bayesian networks*

Bayesian networks can be used to represent the distribution of joint probabilities across the different variables or features describing a classification problem (Jensen and Nielsen, 2007). Such a structure of joint probabilities is inferred using a training dataset, which later can be used to estimate probabilities linking specific features and class values. In these networks a node represents a feature (variable), and the arcs linking the nodes represent statistical dependencies between the features. A network is fully defined by the different conditional probability distributions representing between-feature relationships. Bayesian networks take advantage of the concept of conditional independence to provide more compact statistical model representations (i.e. less complex network structures). If two features are said to be independent given the state of a third variable, such as the diagnostic class, then the two features are said to be 'conditionally independent'.

Bayesian networks can be applied to features characterized by both discrete and continuous numerical variables. The learning of the statistical parameters of a model (i.e. conditional probability distributions) can be implemented by searching for the set of parameters that maximizes the likelihood that the training data are derived from (or fitted to) the model (Jensen and Nielsen, 2007; Needham *et al.*, 2007).

Several algorithms exist to infer the structure of these networks from a training dataset. The main idea is to find a structure of nodes and arcs that maximizes classification criteria given the different samples in the training dataset. The learning algorithm searches, scores, and selects 'good' structures (Needham *et al.*, 2007). Thus, a network construction process can be implemented using different variations of search algorithms, such as greedy search procedures and genetic algorithms (Witten and Frank, 2005).

Networks can also be constructed based on manual or expert-driven approaches, which apply prior knowledge about between-feature relationships relevant to the classification problem under investigation. Once the network structure and corresponding probabilistic model have been specified, 'unseen' or test samples can be used as inputs to make predictions. Several commercial, free and open-source software packages are available to assist in the development and evaluation of Bayesian networks, one of which is introduced by Witten and Frank (2005). Jensen and Nielsen (2007) offer a comprehensive analysis of Bayesian networks including their mathematical foundations. Shorter introductions by Witten and Frank (2005) and Needham *et al.* (2006, 2007) are also recommended.

### *Instance-based learning*

Instance-based learning classifiers include a variety of algorithms based on the idea of processing previous (training) cases and their (correct) classifications to make predictions on new samples (Hastie, Tibshirani and Friedman, 2001). Given a test sample, the goal is to search the training set to retrieve the 'most similar' samples to the test sample. Based on these 'nearest neighbours', a prediction is made based on the classes assigned to the retrieved cases. One prediction approach consists of using the majority class (or a weighted function) of the '$k$-nearest neighbours' to the sample being tested. This category of classifiers is known as $k$-nearest neighbour models. Different types of instance-based learners are available, which differ in the way they estimate similarity between test and training cases, in the procedure employed to define neighbourhoods, and in the class prediction methodology applied (Witten and Frank, 2005). Instance-based learners are also referred to as 'lazy learners' as there is no prediction model construction or learning prior to the processing of test instances.

Table 3.1 summarizes major strengths and limitations of the techniques overviewed above to aid the researcher in making a goal-driven, problem-specific selection.

## 3.4 Feature selection for biomarker discovery

The filtering and selection of predictive features is fundamental to assist in the identification of potential biomarkers. A key reason is that the construction of prediction models using a large number of input features will prevent prediction generalization due to the noise and the low discriminatory power associated with the majority of such features. Of special importance are those techniques used in combination with class-membership labels or specific classification models. Unlike different methods for data visualization or transformation, such as principal component analysis (PCA) (Ringnér, 2008), feature selection techniques preserves the original representation of the input data.

Feature selection in biomarker studies is important not only because it can significantly reduce the dimensionality of the biomarker space, but also because it can offer both quantitative and qualitative insights into the relative (biological or statistical)

**Table 3.1** Examples of strengths and weaknesses of key biomarker data mining techniques

| Technique | Strengths | Limitations |
|---|---|---|
| Decision trees | Interpretation of predictions is intuitive and graphically explained to the user. Fast computing performance when predicting new samples. Different types of data can be processed, including mixtures and missing values. | Large trees or trees with leaves with highly segmented data may produce unreliable predictions. Very sensitive to the size and composition of training datasets. Larger trees may be needed when increasing classification complexity. |
| Logistic regression | Mathematical model easy to interpret in terms of risk scores and coefficients relating markers and responses. | Poor performance when features are correlated. Large datasets may be required for multi-variable prediction models. |
| Artificial neural networks | Powerful modeller of non-linear relationships, noisy and complex classification tasks. | Prediction model or outcomes may be difficult to interpret or explain. Problems involving many inputs may require large amounts of data. Sensitive to selection of learning parameters. Standard models cannot deal with missing data and some of them may be computationally intensive during training. |
| Support vector machines | Powerful prediction performance for complex classification problems. Relatively robust to learning parameter selection. Different types of data can be processed or combined in a single model. | Model implementation difficult to explain or predictions unsuitable for human interpretation. Selection of learning parameters and kernel function can be problematic in some applications. |
| Naïve Bayesian classifier | Simple to implement and interpret. | Poor prediction performance when input features present multiple dependencies. Significant amounts of data are needed to provide accurate estimates of probabilities. |
| Bayesian networks | Prediction performance robust to small perturbation of training data. They can be used to encode background or expert knowledge, and can deal with different data types and missing values. Graphical representation of the model facilitates interpretation of predictions and model. | In applications comprising a large number of features, the automatic inference of network structure and underlying probabilistic model can be computationally intensive and sensitive to data composition and the inference algorithm selected. Prediction performance can be highly sensitive to estimation of prior probabilities and distribution assumptions used to build the model. |

**Table 3.1** (*Continued*)

| Technique | Strengths | Limitations |
|---|---|---|
| Instance-based learning | Relatively robust to dataset size and noise. Models and predictions easy to interpret. | Search and retrieval of cases from the training set may be computationally expensive. Some applications may be very sensitive to the selection of the classification parameters such as the number of k-nearest neighbours and the similarity metric selected. |
| Random forests | Apart from some of the main advantages offered by standard decision trees, it implements embedded feature selection (see next section), show reduced risk of over-fitting, requires minimum user intervention. | Risk of biased feature selection. Feature selection may be sensitive to the choice of relevance metric (Strobl *et al.*, 2007). Feature selection applications can generate multiple solutions, that is feature set instability. In some applications a large number of trees may be required to achieve generalization, and classification performance may be sensitive to the size of the trees implemented (Statnikov, Wang and Aliferis, 2008). |

relevance of specific potential biomarkers in a classification task, such as putative biomarker-specific risk scores or weights. This section overviews feature selection methods for supervised classification applications, that is problems in which the class labels for each sample are known in advance.

A great variety of feature selection techniques and applications have been reported for bioinformatics and biomedical informatics applications. They can be organized on the basis of the search methodology implemented to find relevant features or the set of optimal features for a given application and dataset. The main categories of feature selection algorithms are: filter, wrapper, and embedded techniques.

Filter methods are implemented independently of any classification technique and are used to identify potentially relevant features based on statistical information extracted from the data. Standard statistical testing of multiple hypotheses falls into this category (Chapter 2), as well as different techniques that assess the relevance of features based on relationships observed between the features and the class values under investigation. Wrapper approaches are implemented in combination with a specific classification model. That is, the search and feature selection process is ‘wrapped around’ a classification algorithm. The feature selection process will then depend on the prediction performance obtained by a sub-set of features when used as inputs to a specific classification technique. The search of relevant predictive features can be performed by implementing different deterministic and randomized search algorithms (Saeys, Inza and Larrañaga, 2007). The evaluation of the classification performance based on a sub-set of features can be based on different criteria, such as overall accuracies, AUC values, and so on. Embedded techniques implement search and selection of features as part of the prediction model building process, that is they are integrated into the classifier construction process. This is the case of algorithms based on decision trees, Bayesian approaches and support vector machines.

Examples of filter approaches are multiple-hypotheses testing procedures (Section 3.2) based on the application of parametric tests (e.g. $t$-test, $\chi^2$ test), non-parametric tests (e.g. rank-based tests, permutation-based correction procedures), and information theoretic measures such as mutual information (Cover and Thomas, 1991; Steuer *et al.*, 2002). Filter approaches aim to detect statistically detectable differential patterns between samples derived from different classes or populations.

More advanced approaches consider possible between-feature and feature-class associations or correlations to find optimal subsets of predictive features. It is known that classification performance can be deteriorated by using highly correlated input features. Thus, a well-known feature selection approach consists of finding a sub-set of features with minimal between-feature correlation and maximal correlation between each feature and the class values (Hall, 1999). The correlation-based feature selection (CFS) method is one example, which estimates between-feature and feature-class dependencies (including non-linear correlations) by applying information theoretic concepts (Cover and Thomas, 1991).

The best known approaches to wrapper applications are 'sequential forward selection' and 'sequential backward elimination' (Kohavi and John, 1997), whose search method is deterministic. In the forward selection algorithm, the search process starts with no features followed by the incremental incorporation of new features according to their contribution to the improvement of overall classification performance. The backward elimination process searches for the optimal subset of features by first considering the complete set of features followed by the incremental elimination of features that do not contribute to an improvement in classification performance. An example of randomized feature search in wrapper techniques is the application of evolutionary computation methods (Kohavi and John, 1997; De Jong, 2006) and randomized hill climbing algorithms (Skalak, 1994). Amongst the members of the evolutionary computation category, genetic algorithms can be wrapped around different types of machine learning techniques, such as support vector machines and $k$-nearest neighbours models, to guide the search for an optimum sub-set of features that maximize classification performance based on different CV schemes.

The method of 'shrunken centroids' proposed by Tibshirani *et al.* (2002) is perhaps one of the best known examples of embedded feature selection in gene expression studies. This method, also known as PAM, implements a feature selection procedure that is embedded in a 'nearest-centroid' classification algorithm, which is a member of the family of instance-based learners. Samples in a training dataset are used to estimate the overall and the class-specific 'centroids' in the dataset. A centroid could be defined as a vector encoding the mean values of each feature in the dataset. In the testing mode, a sample is assigned to the class with the nearest centroid to the tested sample. Feature reduction occurs when the distance between the class and the global centroids is reduced by an amount predefined by the user. The stronger this shrinkage, the closer the class centroids will move to the global centroid. When for a given feature the distance between the global and a class distance is reduced to zero, the user is in fact detecting irrelevant or non-differentially expressed features, which become part of the set of features eliminated.

Table 3.2 presents a brief guide on the strengths and limitations of different feature selection approaches. The reader is also referred to reviews published by Guyon and

**Table 3.2** Examples, strengths and limitations of major feature selection techniques relevant to biomarker discovery research

| Type | Strengths | Limitations | Examples |
|---|---|---|---|
| Filter | Computationally inexpensive, easy to implement and adapt to different application domains. Faster and more robust than wrapper approaches. Independent of choice of classifier. Some approaches can model feature relations. | Methods based on standard univariate analysis or multiple-hypotheses testing ignore between-feature or feature-class dependencies. Features selected may be less suitable or powerful when used as inputs to different classifiers. | Multiple-hypotheses testing procedures. Information theoretic approaches (Steuer *et al.*, 2002). Correlation-based feature selection (Hall, 1999). |
| Wrapper | Feature relationships or dependencies can be considered. Domain-specific classification models and performance are taken into account. | Potential over-fitting. Selection bias if learning-testing phases are not properly implemented or isolated. Some approaches, such as those based on random optimization, may be computationally intensive. Selection may be very sensitive to data sampling, that is less robust and unstable than filter approaches. | Sequential forward selection, backward elimination, hill climbing, genetic algorithms (Kohavi and John, 1997). |
| Embedded | Less computationally intensive than wrapper techniques. Feature dependencies are taken into account. Fully integrated into a specific classification learning model. | Constrained to classification model applied. In the presence of insufficient data or relatively unnecessary model complexity, there may be a significant risk of over-fitting. | Decision trees, Bayesian methods, selection based on support vector machines, PAM (Tibshirani *et al.*, 2002; Guyon and Elisseeff, 2003). |

Elisseeff (2003), which provide a detailed analysis about algorithm implementation, and by Saeys, Inza and Larrañaga (2007), which discusses some of their applications in bioinformatics.

The relation between the number of optimal features and sample size represents another major question in biomarker classification problems. As in the case of sample size estimation in traditional hypothesis testing (Chapter 2), with univariate or multivariate models, this relation is complex to estimate and depends on the scope and context of the application. Moreover, ‘universal’ estimation tools could not be feasible because optimal sample and feature set sizes will depend on the classification technique studied and the statistical distribution of the features and classes. However, it is important to have an idea about potential optimal relations using, for example, comprehensive knowledge-based evidence or approximations based on previous studies. In the case of microarray data studies, Page *et al.* (2006) offer a Web-based system to aid researchers in the estimation of optimal sample sizes based on published research. But this database does not explicitly consider the influence of the number of features and types of classification models. Hua *et al.* (2005) performed large-scale analyses of feature-sample size estimations using synthetic and real microarray datasets, as well as different classification techniques of diverse complexity. The main outcome of their research was a collection of 3D surface maps that depict the relation between classification error, number of samples and number of features for different classifiers, learning parameters, datasets and data constraints (e.g. between-feature correlations). Although extreme caution should always be exercised in the presence of ‘rules-of-thumb’ or ‘universal’ solutions, this type of resources can be valuable tools to assist researchers in the understanding of problem requirements and identification of potential approaches.

## 3.5 Critical design and interpretation factors

The problem of selection bias (Ambroise and McLachlan, 2002) has been reported as a major pitfall in recent advances in biomarker discovery and ‘omic’ data analysis when wrapper or embedded feature selection is performed. This problem refers to an overestimation of the prediction performance when the model learning phase is not completely separated from the testing phase. A typical scenario is the classification-based selection of predictive features (e.g. a wrapper approach) using all the data available, followed by the evaluation of the best classifier (and the most relevant features) through CV.

In order to obtain unbiased and accurate estimates of prediction performance the data used to select features should not be used to test the resulting classifier. That is, feature selection based on the application of a specific classification model should be considered as part of the model building process. In practice, the relative lack of samples available for building, testing and independently validating prediction models may make the prevention of selection bias a difficult task. Nevertheless, it is important to correct this bias by implementing a classification CV procedure that is external to or independent of the biomarker selection process (Ambroise and McLachlan, 2002). This also means that the (wrapper or embedded) feature selection should be implemented during the training of the classification model, at each stage of the CV procedure.

This concern for reduction of bias also applies to situations in which the prediction model requires the selection of optimal learning parameters, which are specific to the model, through CV. Examples of such learning parameters are the shrinkage parameter, $\Delta$, in the PAM method (Tibshirani *et al.*, 2002), or the complexity parameter, $C$, in a support vector machine. Varma and Simon (2006), for instance, demonstrated that biased estimates of prediction performance can be obtained when incorrectly using CV on a model whose learning parameters were actually obtained (or tuned) using CV. The recommendation is to perform the search for optimal learning parameters as part of each CV stage.

An example of this approach, in the case of the shrunken centroids method, is to perform LOOCV with nested 10-fold CV (in each training partition) to select learning parameters. Given $N$ samples, one sample is left out and the shrinkage parameter, $\Delta$, is selected on the remaining $N-1$ samples based on 10-fold CV, that is the optimum $\Delta$ is selected if it generates the maximum classification accuracy based on the 10-fold CV. The resulting classifier is tested on the left-out sample, and the process is repeated until the LOOCV is completed (Varma and Simon, 2006).

The prevention of major drawbacks and pitfalls in the design and interpretation of biomarker-based prediction models depends on a solid understanding of the research context and its limitations, as well as on a clear definition of research prediction goals, and the potential relevance of model inputs and their associations. Another key factor is the specification of evaluation and outcome acceptability criteria prior to the data acquisition and prediction model investigation.

A multi-disciplinary approach is important to answer questions regarding the type of representations of models and predictions, and the types and domain-specific meaning of prediction uncertainty indicators, for example prediction quality error indicators, confidence measures. In some applications predictive accuracy may be the most important factor to assess the potential relevance or validity of biomarker model. In other cases, its interpretability and robustness to data availability and definition of learning parameters may be more crucial. For instance, given that two prediction models offer similar classification performances, key questions are: which model is easier to understand? Which one allows the user to have a better assessment of its reliability? (König *et al.*, 2008; Bellazzi and Zupan, 2008). Moreover, different modelling or classification approaches may represent valid solutions based on different criteria. In this case one may give priority to relatively less complex solutions, or to those based on assumptions relatively easier to justify or verify, such as assumptions relating to statistical distributions of the data or potential dependencies between features.

Recent critical reviews of the literature relevant to biomarker discovery based on microarray data have highlighted major flaws and limitations (Dupuy and Simon, 2007). Despite recent advances, the lack of or inadequate strategies for multiple-testing still deserves more careful consideration (see Chapter 2). Or at least there is still a need to remind researchers of the importance of clearly stating null hypotheses and correction strategies in multiple-testing applications. Dupuy and Simon (2007) also consider that, at least until 2007, many researchers were still making spurious claims about relationships between data clusters and clinical outcomes due to biased estimations of feature-class associations. The problem of overestimating predictive capability is also found in supervised classification applications through selection bias as pointed out above. A detailed description of these flaws and recommendations to address them, with an emphasis on microarray-based cancer biomarker studies, was published by Dupuy and Simon (2007).

### *More on evaluation*

In order to prove the clinical usefulness and validity of newly discovered biomarkers, it is important to report its predictive performance in comparison to traditional prediction models. The latter are typically based on markers already available in routine clinical practice. This also involves comparisons against models that combine traditional and the proposed biomarkers. It has been recommended that these evaluations should be implemented on datasets that are independent from the datasets used during the ‘novel biomarker’ discovery phase. Another important evaluation criterion is that the traditional biomarkers (alone) should produce prediction performances consistent with their known clinical performance or those obtained in previous research. Furthermore, correlations between traditional and proposed biomarkers should be minimized.

Which indicators are more informative? It depends on the goals and potential error costs of the application. For example, in an application involving the screening of an uncommon disease (or complication) in asymptomatic individuals, a high specificity may be preferred. In this case the cost of incorrectly classifying a healthy patient will be higher than the cost of missing a (rare disease) positive case. Moreover, this type of application will also include different diagnosis confirmation tests. A different focus would be required in the diagnosis of a relatively more common, life-threatening disease. In this situation, the cost of missing a true positive prediction outweighs the cost of mislabelling a negative case (i.e. predicted as diseased when the patient is actually healthy). Therefore, in this case biomarkers or prediction models capable of generating higher sensitivities should be preferred.

Several quantitative indicators may also be applied to prediction models that produce numerical outcomes (e.g. risk scores). These indicators aim to assess the quality of the predicted values for the prediction of clinical events (expected events) in comparison to the actual events observed in a population. Different probabilistic approaches can be used to estimate predictive quality on the basis of uncertainty, confidence or reliability, as in the case of Bayesian classifiers. In some traditional, single-biomarker applications, higher (or lower) biomarker values are linked to increased (or decreased) risk levels. Such associations are quantified, for example, using hazards and odds ratios (derived from survival analysis), as well as probabilistic estimates of biomarker-disease association strength. Caution should be exercised when interpreting these indicators in studies involving one biomarker only. For instance, a high hazard ratio for a single biomarker-disease association does not guarantee a good classification performance for different predictive thresholds. This is because the biomarker value distributions of individuals from different clinical groups (control vs. disease) will always have some overlap. Therefore, significant associations between (single) biomarkers and clinical condition may be used as evidence to define risk factors, but not to suggest strong prediction capability by itself, even after accounting for confounding factors. But independently of the required outcome representation, prediction technique, predictive feature combination method or evaluation technique, a key challenge is to show how the new or proposed prediction model can outperform conventional or standard prediction models.

The selection of evaluation metrics and procedures is as critical as the selection of patient samples and classification models. Chapter 2 introduced some of these indicators, together with key strengths and challenges present in this area. Although, the adaptation of known or 'generic' predictive accuracy indicators to specific purposes in different areas has been proposed (Swets, 1988), it is important to consider, above all, the meaning and possible interpretations that such methods may represent in a specific context or application. In the case of widely used techniques, such as the ROC curves and AUC, it is important to recognize their limitations and mathematical meaning (Chapter 2) to prevent misuses and biased comparisons between prediction models (Cook, 2007, 2008). Moreover, when comparing different prediction models or biomarkers, the demonstration of the statistical 'significance' of their differences or relations is not sufficient to prove the validity of a new biomarker model (Pencina *et al.*, 2008).

Despite its limitations and misinterpretations in the clinical context (Chapter 2), ROC curves and AUC will continue to be considered as valuable tools or criteria to assess discrimination capability in biomarker studies. This could be the case especially in studies with extreme (very large or very small) differences in performance between prediction models (Pencina *et al.*, 2008), or with small datasets (Pepe *et al.*, 2008a, 2008b). Nevertheless, it is important to consider other indicators, for example those based on a closer look at sensitivity/specificity values (Pepe, Feng and Gu, 2007; Pepe, Janes and Gu, 2007; Pepe *et al.*, 2008a, 2008b), which reflect alternative predictive quality properties and context-specific requirements. Moreover, it has been shown that standard methods for prediction performance evaluation can be both biased and inaccurate (Wood, Visscher and Mengersen, 2007). Therefore, the calculation and interpretation of different indicators in the light of context-specific study goals and requirements is recommended.

### *The top of the class*

The selection of techniques for implementing prediction models is context-dependent, and involves the analysis of previous research in related application domains. As pointed out above, different techniques may exhibit shared advantages and limitations, which may guide the selection of potential solutions. Moreover, it is important to implement benchmark or baseline techniques to support the evaluation of new or alternative solutions, as well as their application to reference biomarkers, especially those currently applied in the clinical environment. Comparative analyses commonly require the application of known reference algorithms, such as naïve Bayesian, or more advanced approaches, such as support vector machines and random forests. The latter options are currently known as two of the most powerful classification models across different application areas (Statnikov, Wang and Aliferis, 2008).

Recent empirical evaluations using microarray biomarkers in cancer classification showed that support vector machines can outperform random forests in different design and application settings, including those in which these techniques are used for automated biomarker selection (Statnikov, Wang and Aliferis, 2008). On the other hand, other studies have shown that in multiple-class classification problems, advanced models, such as support vector machines, may display a poorer performance depending on critical design choices. Support vector machines, which were originally designed to deal with two-class problems, can be combined to form ensembles of classifiers that can be used to predict multiple classes according to binary classification schemes. This strategy may comprise the combination of multiple one-versus-one or one-versus-others individual classifiers. Comprehensive comparative evaluations of different multi-class methodologies for the molecular classification of cancers has shown, for instance, that the predictive performance of multi-class support vector machine models may be sensitive to the selection of the two-class model combination scheme (Statnikov *et al.*, 2005). Diverse or inconsistent predictive performances might also be obtained with relative minor variations in learning conditions, such as the number of samples, feature selection method and composition of the input features, even for the same classification model (Pirooznia *et al.*, 2008).

Apart from model selection bias, the problem of sample collection bias may represent another obstacle to obtaining accurate estimates of prediction performance and to supporting independent validations. Ransohoff (2005) and Ressom *et al.* (2008), for instance, have reported that bias can occur if the experimental and control samples are obtained, stored and processed using different methodologies or experimental protocols. Such artefacts or analytical variability may explain the 'significant' differences found in statistical analysis.

It is also important to stress that biomarker data mining and prediction model implementation should be seen as incremental, iterative and interactive processes. Prediction model design, including classification implementation and/or feature selection, is typically preceded by multiple steps of data pre-processing, visualization and exploration (Azuaje and Dopazo, 2005). The latter two stages may involve different unsupervised analytical tasks, such clustering-based visualizations. Recent advances include several techniques based on matrix decomposition (Schachtner *et al.*, 2008), such as independent component analysis and non-negative matrix factorization. These techniques may also be applied to extract predictive features, which can represent inputs to subsequent supervised biomarker data mining tasks (Pascual-Montano *et al.*, 2006). It is very likely that the combination of different exploratory data analysis tools (Witten and Frank, 2005; Montaner *et al.*, 2006; Mejía-Roa *et al.*, 2008) and prediction model implementation engines will continue to be a driving force in biomarker discovery. However, there is still a need to develop integrated, user-friendly infrastructures and solutions tailored to biomarker discovery (Chapter 9).

# 4 An introduction to the discovery and analysis of genotype-phenotype associations

This chapter presents an introduction to genomic data and approaches to biomarker discovery: DNA variation markers and genome-wide association analysis. It will introduce fundamental concepts, such as linkage disequilibrium, the Hardy-Weinberg equilibrium and genetic interactions. The characteristics of the data and technical requirements will be discussed. This will include discussions on recent advances in cardiovascular and cancer research, concentrating on: (a) biomedical findings and clinical applications, (b) statistical and data mining methodologies applied, and (c) strengths and limitations. Chapter 9 will offer additional information on databases and software relevant to genome-wide association studies.

## 4.1 Introduction: sources of genomic variation

Alleles and genotypes represented in statistically detectable high rates in a population sample may be seen as factors that confer greater susceptibility to a particular disease or clinical response. Moreover, these *loci* may be strongly associated with other (phenotypically-neutral) loci, that is in linkage disequilibrium, which may also be over-represented in the population sample. Thus, one of the main objectives of genotype-phenotype association studies is to identify statistically detectable differences between phenotype-specific groups of individuals (e.g. between cases and controls) on the basis of their genotypes (Li, 2008). In recent years, a significant amount of investigations on

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

putative (genomic variation) biomarkers have been reported. For example, as of January 2009, more than 200 publications reporting more than 900 SNPs-phenotype associations in more than 70 common diseases had been published (El-Omar, Ng and Hold, 2008; Hindorff *et al.*, 2008).

The main types of genotype-phenotype association studies are: *candidate-gene* and *genome-wide association* studies. The former are hypothesis-driven studies involving genes for which there is evidence of a possible association with a disease or clinical outcome. This commonly requires the sequencing of the gene in case-control groups and the search for variants that differentiate these groups. Genome-wide association studies are based on the unbiased investigation of variants across most of the genome.

Potential genomic variation-based biomarkers can be initially predicted by finding strong genotype-phenotype associations in a specific population sample. Two important sources of DNA variation are SNPs and gene 'copy number variation' (CNV). Figure 4.1 illustrates these concepts. SNPs are sequence variants involving a single nucleotide. SNPs can affect gene expression and protein function by altering not only protein-encoding regions, but also non-coding (intronic or regulatory) regions. Variations detected in non-coding areas may have important functions in the development of biomedical phenotypes. Such variations could actually alter gene expression and splicing (Hardy and Singleton, 2008; Altshuler, Daly and Lander, 2008). This motivates additional research, for example, on the potential causative roles of variations found in highly-conserved non-coding regions (Hirschhorn and Daly, 2005). This may be particularly difficult because some of these regulatory sequences may be found far from the coding regions.

A DNA locus may be defined as polymorphic if at least two variants (e.g. alleles A and G) can be found in the locus, and if the frequency of the most common variant is less than 99% (Landegren, Nilsson and Kwok, 1998). Examples of SNPs-disease associations include a variety of SNPs connected to susceptibility to myocardial infarction and coronary artery disease, such as variants found in the genes ALOX5AP, LTA4 and LGALS2 (Topol *et al.*, 2006). Several polymorphisms in Toll-like receptor genes have been documented and associated with the risk of gastric and prostate cancers (El-Omar, Ng and Hold, 2008). An increasing number of replicated associations in different medical areas have been reported (Hindorff *et al.*, 2009).


**Figure 4.1** Major sources of genomic variation: SNPs and CNVs

CNVs comprise genomic variations in which blocks of DNA are missing or duplicated (Figure 4.1). Thousands of CNVs have been reported in different populations, and international organizations and consortia are cataloguing patterns of CNVs in different diseases. Different potentially-significant associations between CNVs and disease have been reported in several cancers, schizophrenia, autism, body mass index, Chron's disease and retinoblastoma (Couzin, 2008). Despite these advances there are still many questions unanswered about the origin, effect and medical relevance of such associations. For example, are these CNVs inherited or spontaneous? Are they actually connected to other genomic variables? Moreover, recent efforts, such as those implemented by the Wellcome Trust Case Control Consortium (Couzin, 2008), have found that the key to understanding the connection between CNVs and disease may not entirely lie in the calculation of differential representations of the same CNVs between disease and control groups, but in the identification of disease-specific CNVs. This would suggest that efforts should go beyond the estimation of the number of common CNVs in healthy and disease groups, and that more attention should be given to the problem of finding out where these CNVs actually occur in each group.

Advances in genotyping technologies, the availability of public databases and the development of the International HapMap Project (The International HapMap Consortium, 2005, 2007) have facilitated the implementation of genome-wide association studies. Genome-wide association studies have so far mainly focused on the identification of SNPs-phenotype associations involving common variants at the SNPs (Donnelly, 2008). This is because the power to predict associations is reduced by the frequency of the less common variant at a particular SNP, that is rare variants are more difficult to detect. Apart from this limitation, rare variants are difficult to detect because genomic-variation information encoded in databases and tagging approaches (Section 4.2) have focused on more common variants (Hirschhorn and Daly, 2005). Moreover, the majority of SNPs-phenotype associations discovered to date are not likely to represent disease causative or functional mutations (Altshuler, Daly and Lander, 2008). Instead, they may only encode a 'proxis' for the actual phenotype driver or causal agent. The next section discusses the fundamental concepts behind these relationships. A significant proportion of the investigations reported to date have focused on genomic variants found in protein coding regions. The challenge is that many of the SNPs with disease causative properties may not be found in protein-encoding regions of the genome, that is they effect function at the gene regulation level.

Genotype-phenotype associations can be used to build risk assessment and phenotype classification models. Different data encoding the relevant genomic variations can be applied as inputs to these models. Moreover, different risk assessment, treatment selection and classification models can be implemented by combining information from genotype-phenotype associations with environmental and traditional risk factors. In this case, one of the challenges for bioinformatics is to find effective and meaningful ways to aggregate these information types based on existing and novel statistical and machine learning methodologies (Chapter 3). It has been suggested that risk scores and classification power of models based on recent association studies may be underestimated because most causal variants have not been identified yet (Donnelly, 2008). On the other hand, one deals with the problem of making sense of a great variety of weak genotype-phenotype associations (e.g. with odd ratios $<2$). Therefore, independently of the identification of causal associations, another important challenge is to integrate such 'weak biomarkers' to improve both our understanding and prediction capability in complex diseases (Loscalzo, 2007).

## 4.2 Fundamental biological and statistical concepts

### *Hardy-Weinberg equilibrium*

A genomic locus is in Hardy-Weinberg equilibrium (HWE) when the two alleles in an individual have been randomly acquired, that is the two alleles are statistically independent. HWE is expected to be observed in populations without a history of significant migration, ethnic admixture or inbreeding. For a particular SNP, deviations from the HWE can be detected by statistically testing the hypothesis that the population is in HWE. For dominant and recessive alleles '*A*' and '*a*' respectively, a population meets the conditions of HWE if the genotype frequencies are: $p^2$ (for genotype *AA*), $2pq$ (for genotype *Aa*), and $q^2$ (for genotype *aa*). Where $p$ is the frequency of *A* and $q$ is frequency of *a*. These frequencies define the 'expected' genotype frequencies under HWE. Thus, the HWE hypothesis can be tested by comparing the genotype frequencies observed, $O$, in the population sample against the expected genotype frequencies, $E$. This can be done by applying hypothesis testing procedures for categorical data (Chapter 2), such as the $\chi^2$ (Chi-2) and Fisher's exact tests. The $\chi^2$ statistic, for example, is calculated as: $\chi^2 = \sum_i \frac{(O_i - E_i)}{E_i}$, with $i: 1, 2, 3$, with 1 degree of freedom. The HWE hypothesis is rejected if the resulting statistic is greater than the value required for a given significance level. That is, a departure from HWE is detected if the corresponding $P$ value is below a significance level, for example $P = 0.001$. These analyses may be supplemented by the graphical visualization of the $P$ values obtained from all the SNPs investigated. One option is to use log quantile-quantile plots of the $P$ values, in which deviations from the diagonal line, $y = x$, highlight loci with departures from the HWE (Balding, 2006).

Departures from HWE (i.e. HWD) in random samples have been commonly used as indicators of errors in experimental work: errors in assays, genotypes or data acquisition. Hosking *et al.* (2004), for example, found associations between genotyping errors, as well as the detection of non-specific SNPs, and deviation from HWE. On the other hand, other studies have shown that genotyping errors may not generate significant departures from HWE (Cox and Kraft, 2006), and indicate that the most effective method to detect these errors is the verification of genotypes by re-sequencing or independent genotyping.

Deviations from HWE have also been suggested as indicators of inbreeding, selection and population stratification (Balding, 2006). As an approach to data quality assessment, researchers typically exclude SNPs from subsequent analysis, when they deviate from HWE at a significance level around 0.0001 in the control group (Ziegler, König and Thompson, 2008). For example, in a recent genome-wide study that linked six new loci to cholesterol and triglycerides (Kathiresan *et al.*, 2008), a significance level of 0.001 was used to exclude SNPs from analyses.

On the other hand, researchers have suggested that a locus in HWD (i.e. departure from HWE) may represent a marker of disease susceptibility heterogeneity, which could be in linkage disequilibrium with other susceptibility loci (Nielsen *et al.*, 1998). Thus, tests reporting departures from HWE at a marker locus may offer useful evidence in the search of key biomarkers.

Also, as Nielsen *et al.* (1998) explained, HWD should not be expected to be found in disease models in which the alleles act in a multiplicative way to increase susceptibility. However, a greater amount of HWD should be expected if the effects of the alleles deviate from multiplicative interactions. Wittke-Thompson, Pluzhnikov and Cox (2005) have discussed the problem of distinguishing genotyping errors from other possible causes of HWD, and concluded that significant departures from HWE can be expected in relatively small samples of patients.

### Linkage disequilibrium and haplotypes

Neighbouring genomic variants or alleles are often correlated because of their shared evolutionary history, that is they have been passed from generation to generation in a common block of DNA. Such a correlation means that it would be possible to infer an allele at a particular SNP based on the allele observed at a neighbouring SNP (Altshuler, Daly and Lander, 2008). This correlation is known as 'linkage disequilibrium'. Thus, in the presence of linkage disequilibrium there is no need to genotype all variants (including the causal variant) to detect potentially relevant associations. This allows researchers to reduce the size of their genome-wide association studies. If in subsequent investigations (e.g. validations) more SNPs are typed in the genomic region of interest, then it would be possible, in principle, to detect the potential causal SNP by finding the SNP showing the strongest association with the phenotype studied (Donnelly, 2008).

Figure 4.2 shows snapshots obtained from a graphical analysis of linkage disequilibrium in chromosomes 20. It illustrates correlations between different SNPs found in a specific region of this chromosome. The colour-coded triangle shown is a correlation matrix that links the SNPs found in this region, with darker cells highlighting strong correlations between pairs of SNPs. Figure 4.3 presents the outcomes of a similar analysis focused on the gene MMP1 on Chromosome 11. In this example the figure highlights a block of strongly correlated SNPs.

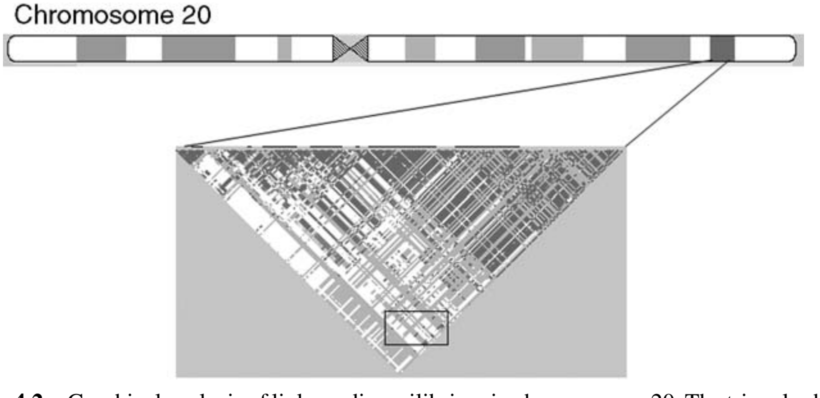


**Figure 4.2** Graphical analysis of linkage disequilibrium in chromosomes 20. The triangle shown is a correlation matrix that links the SNPs found in this region, with darker cells highlighting strong correlations between pairs of SNPs. Analysis implemented with Haploview (Barrett *et al.*, 2005)

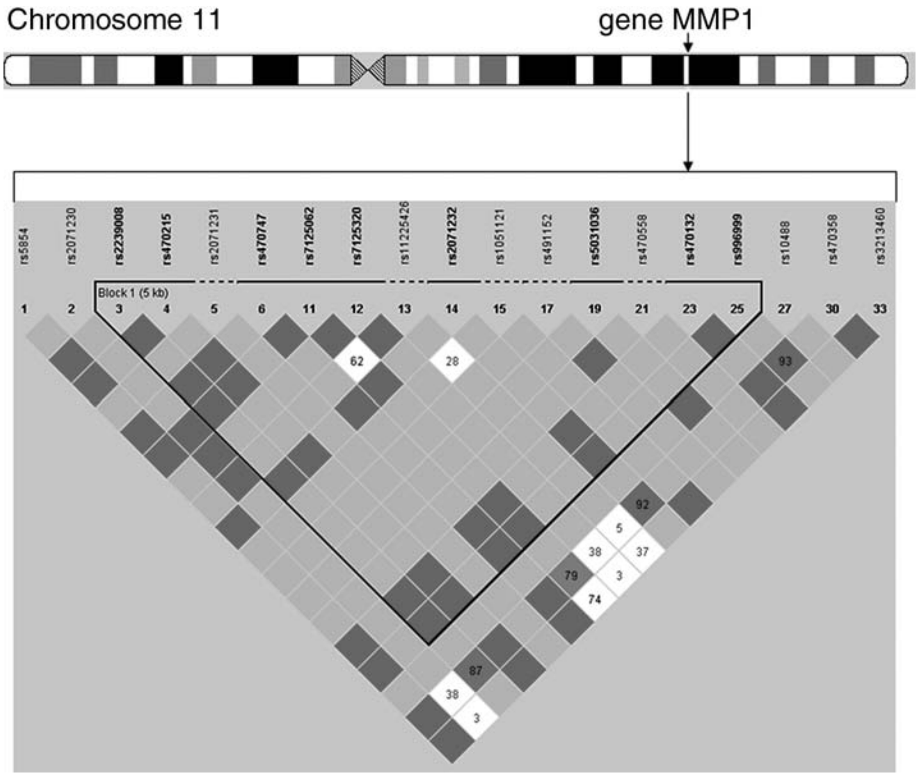


**Figure 4.3** Graphical analysis of linkage disequilibrium focused on gene MMP1 on Chromosome 11. Analysis implemented with Haploview (Barrett *et al.*, 2005)

Groups of polymorphisms in strong linkage disequilibrium can be used to define 'haplotypes' (Altshuler, Daly and Lander, 2008), which represent most of the genomic variation within a specific chromosomal region. Haplotypes comprise DNA regions located between recombination hotspots, that is haplotypes consist of stretches of DNA that tend to be inherited together from generation to generation (Musunuru and Kathiresan, 2008).

Most of the genome can be characterized by regions of strong linkage disequilibrium and most chromosomes include one haplotype (Hirschhorn and Daly, 2005). Thus, researchers can in principle identify all relevant SNPs-phenotype associations based on the identification of a few SNPs within a haplotype. This assumption and the estimations of the number of haplotypes have allowed researchers to argue that a 'genome-wide' scan of 500 000 SNPs can be used to cover more than 90% of SNPs-related variation in the human genome (Loscalzo, 2007; Hardy and Singleton, 2008) in non-African populations (Kruglyak, 2008). Haplotypes can also be used as inputs to phenotype prediction models based on statistical and machine learning (Malovini *et al.*, 2009). Figure 4.4 illustrates a hypothetical example of the identification of three haplotypes, which are detected in a sample of 10 DNA sequences and 3 SNPs.

Two of the best known measures of linkage disequilibrium, for two-locus haplotype data, are the $D'$ and $r^2$ measures (Li, 2008). These measures are based on the observed

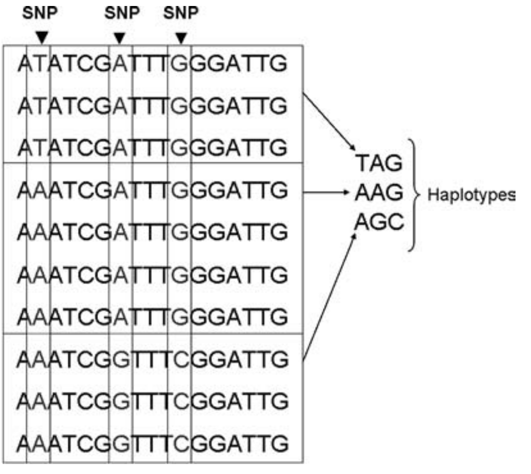


**Figure 4.4** Illustration of the concept of haplotypes. Three hypothetical haplotypes are identified in a sample of 10 DNA sequences and 3 SNPs

frequencies of the alleles from the two markers studied, and can be implemented by different software tools, such as Haploview (Barrett *et al.*, 2005) and PLINK (Purcell *et al.*, 2007). It has been suggested that D′ can overestimate the amount of linkage disequilibrium even in the presence of a very rare allele (Balding, 2006). Because these are two-locus measures, researchers have to adapt them to estimate linkage disequilibrium over specific chromosomal regions. One typical approach is to calculate the average of the different (pairwise) measures over the region of interest. The colour-coded visualization introduced above has become a standard methodology to assess within-region relationships and to support the detection of haplotype blocks (Figures 4.2 and 4.3).

The estimation of linkage disequilibrium allows researchers to select a small number of SNPs as representatives of the genomic variation encoded in the genome or in a larger set of SNPs under investigation. This task is referred to as 'tagging'. This is an important task to aid in the reduction of the amount of SNPs to be typed in a large-scale association study. Similarly, it can also assist researchers in focusing their analyses on a relatively small subset of SNPs already typed in a population sample, or in the imputation of additional SNPs without having to perform additional genotyping (Musunuru and Kathiresan, 2008). One typical tagging approach consists of the analysis of the $r^2$ values for all the pairs of SNPs. From each pair of SNPs, one of the SNPs is excluded from subsequent analyses if, for instance, the pair shows $r^2 > 0.9$.

With the development of next-generation sequencing technologies linkage disequilibrium and haplotype analyses will be eventually replaced by the large-scale genotyping of all known common SNPs (Kruglyak, 2008). New genotyping technologies will also address some of the limitations of analyses based on the assumption of linkage disequilibrium and haplotypes. For example, the phase 2 of the HapMap project found that the assumption of indirect association-mapping based on linkage disequilibrium does not hold for approximately 1% of all SNPs (The International HapMap Consortium, 2007). This means that such SNPs are 'untaggable' and require direct genotyping in order to explore phenotypic associations.

## 4.3 Multi-stage case-control analysis

The implementation of multi-stage association studies is motivated by the need to: (a) reduce the rate of false positive predictions, (b) reduce the rate of false negative predictions, and (c) reduce the cost of genotyping large population samples. The first phase typically involves the genotyping of a relatively large number of SNPs in a relatively small population sample. This phase aims to filter as many potentially irrelevant associations as possible, while trying to maintain an adequate level of statistical power (Hirschhorn and Daly, 2005). This is achieved by using relatively modest (or relaxed) statistical significance levels when testing the multiple hypotheses (see below). Subsequent phases are implemented to validate the resulting associations, that is the markers that passed the first stage are analyzed in independent population samples. In these phases more individuals are incorporated into the case-control groups and more stringent significance levels are selected. Common practices include the application of different genotyping techniques and the involvement of independent research groups in the different phases. This is useful to aid in the reduction of spurious associations and to improve the scientific credibility of the validations (Hirschhorn and Daly, 2005). In the initial stage of a two-stage (or multiple-stage) association analyses, cross-validation procedures (Chapter 3) could be useful tools to support the assessment of the potential relevance and validity of the predicted associations. This has been suggested as a valid alternative to independent validation in the exploratory phases of a genome-wide association study (Loscalzo, 2007).

An example of a two-stage association study can be illustrated by a project that aimed to detect SNPs associated with breast cancer in the Spanish and Finnish populations (Milne *et al.*, 2006). In the first phase, the genotype frequencies of more than 640 SNPs in 111 genes were analyzed in 864 breast cancer cases versus 845 control individuals. This initial set of SNPs was selected by focusing on known cancer-related genes and variants selected by tagging. The first stage reported 10 SNPs as significantly differentially observed in the two groups with a (nominal) $P$ value below 0.01, without further corrections for multiple testing. In stage 2, the SNPs derived from stage 1 were analyzed in larger case and control groups. Out of these 10 SNPs, one SNP (on intron 1 of ERCC4) was associated with breast cancer protection after correcting for multiple testing. This SNP reported a $P = 0.04$ after Bonferroni correction. These results were also supported by a permutation-based correction procedure.

Figure 4.5 illustrates a hypothetical two-stage association study, with typical analytical steps and outputs. In this figure, the amount of data (individuals and SNPs) included in each stage is graphically reflected on the size of the data symbols shown on the right-hand side. Note that additional stages can be implemented between these stages.

## 4.4 SNPs data analysis: additional concepts, approaches and applications

The basic approach to finding genotype-phenotype associations is to apply statistical hypothesis-testing procedures. The null hypothesis to be tested is that there is no detectable difference between two populations, such as two groups of patients belonging to two diagnostic classes, on the basis of the genotype frequencies (i.e. genotype

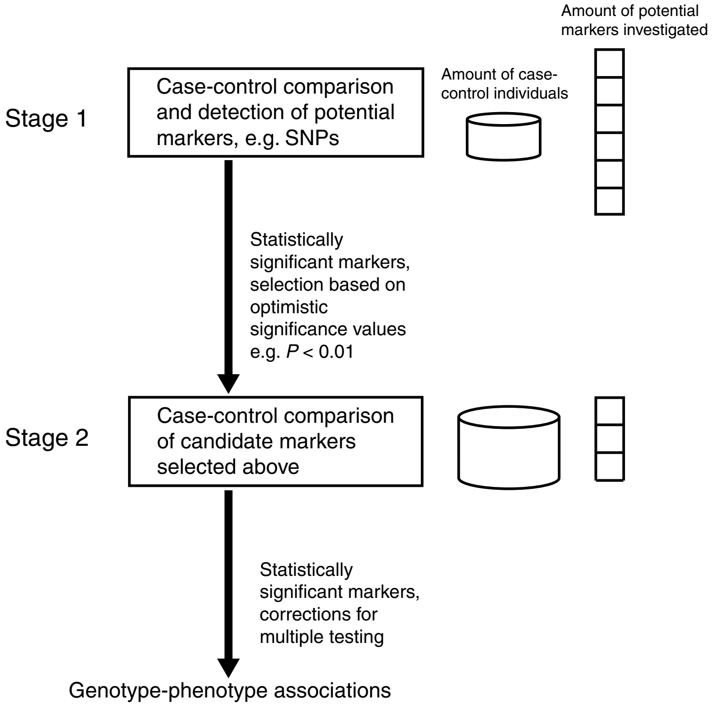


**Figure 4.5** Overview of main steps and outputs in a typical two-stage association study. Hypothetical example comprising the comparison of case and control groups on the basis of SNPs. The amount of data (individuals and SNPs) studied in each stage is graphically reflected on the size of the data symbols shown on the right-hand side

proportions) observed in each group. As explained in Chapter 2, this may be done on a $2 \times 2$ contingency matrix, with the 2 groups vs. 2 genotypes (homozygous and heterozygous genotypes); or on a $2 \times 3$ contingency matrix if the comparison requires the inclusion of the 2 homozygous genotypes separately. The Chi-2 and Fisher's exact tests are the typical tests applied in this task (Chapter 2). In one example, associations between a polymorphism in the MHC2TA gene and cardiovascular mortality after myocardial infarction were detected by applying the Chi-2 test to patient and control groups consisting of thousands of individuals (Lindholm *et al.*, 2006). The Fisher's exact test is recommended when the expected frequency of any of the genotypes studied is smaller than five. Odds-ratios (Chapter 2) are also commonly reported to illustrate differences between groups on the basis of their genotype frequencies.

The application of statistical analysis, including the selection of between-group comparison procedure, depends on the genetic model being tested: Dominant, recessive or additive models. Given alleles '*A*' and '*a*', in the dominant model one is interested in comparing the (dominant) homozygous genotype (*AA*) versus the other genotypes (*Aa* and *aa*). In the recessive model, one compares the (recessive) heterozygous genotype (*aa*) against the genotypes *AA* and *Aa*. The additive model assumes the combined, linear effect of the three genotypes (Ziegler, König and Thompson, 2008).

Apart from statistical hypothesis testing, more advanced genotype-phenotype association modelling can be implemented based on classification models, in which the inputs represent the different genotypes for a particular SNP and the phenotype is the class to be predicted (e.g. diagnostic class). Different statistical and machine learning techniques, such as logistic regression and support vector machines (Chapter 3), can be applied for this purpose. These techniques can also be adapted to association studies involving phenotypes measured on a continuous numerical scale. For example, recent studies that linked new loci to cholesterol (Kathiresan *et al.*, 2008; Willer *et al.*, 2008) applied multivariable linear regression (Chapter 3) models in which lipid concentration values represented the phenotype or outcome to be estimated.

Multivariable statistical and machine learning models allow the incorporation of genotype information derived from multiple SNPs, as well as environmental exposure factors and traditional biomarkers. Moreover, potential 'epistatic interactions' may be explored, as further discussed in Section 4.6. However, note that multiple-SNPs models may have little impact on (or deteriorate) phenotype predictions if there is a single causal variant. Tagging and different feature selection procedures (Chapter 3) are also useful to improve the prediction performance of multiple-SNPs models. The former can be applied to problems with a large number of SNPs. Feature selection (Chapter 3) is recommended to reduce the number of highly-correlated SNPs. Because haplotypes capture the correlation structure of SNPs, they can also be used to model the inputs to phenotype classification models (see below).

Apart from genotyping errors, such technical artefacts and inadequate standardization of experimental protocols, population stratification represents an important source of confounding in association studies. Population stratification refers to the existence of different sub-groups within a population, which differ in terms of disease prevalence or other genotype-phenotype distinguishing features. The problem posed is that (false) positive predictions (i.e. differences, associations) may be detected between case-control groups, which are actually explained by the presence of stratification. The definition of well-matched case-control groups is an effective strategy to prevent large-scale stratification (Hirschhorn and Daly, 2005).

Different techniques have been proposed to detect stratification based on statistical hypothesis testing procedures and comparisons involving genomic control procedures (Balding, 2006). One such approach is based on 'null SNPs'. A null SNP is a SNP with no true association with the phenotype investigated. A genomic control procedure can be based on the application of the Armitage test (Glantz, 2001), which is another method for comparing categorical data (Chapter 2), to each of the null SNPs. The parameter $\lambda$ is defined as the empirical median of the obtained Armitage test statistics divided by its expected value, under the Chi-2 distribution with 1 degree of freedom. If $\lambda > 1$, then population stratification is likely to be present (Balding, 2006). This is because one should expect no, or few, null SNPs associated with the phenotype.

Population stratification may also be identified by implementing visualization-based exploratory analysis of the null SNPs. This can be done, for example, by applying unsupervised learning techniques or principal component analysis to identify subgroups of similar individuals, or clusters, within the population sample. It has been suggested that these approaches may be more reliable than simply using information on geographical origin or ethnicity as criteria for inferring stratification (Balding, 2006; Altshuler, Daly and Lander, 2008).

The problem of missing genotypes or incomplete data in SNPs data analysis is not uncommon. Typically, SNPs are excluded if their total missing frequency across the groups studied is greater than 2% (Ziegler, König and Thompson, 2008). Nevertheless, different algorithms are available to implement missing data estimation, such as those based on likelihood estimates for single imputations, multiple imputation based on random sampling, regression models and nearest neighbour methodologies (Balding, 2006). For instance, missing data can be estimated based on the genotypes observed at neighbouring SNPs. Another approach is to make an estimation based on the genotype observed in another individual, whose genotype is similar to the neighbouring genotype in the individual with the missing value.

The size of the human genome and the diversity of genomic variants make the problem of multiple-hypotheses testing a critical issue. Such a complexity means that in theory any SNP is unlikely to be associated with any given phenotype. The concepts and techniques introduced in Chapter 2 can be applied to this area, but special attention should be given to the satisfaction of underlying assumptions and application constraints. For instance, the Bonferroni correction could be a very conservative approach when dealing with highly-correlated SNPs. This is because the between-variable statistical independence assumption of the Bonferroni correction is not satisfied. Therefore, the FDR and permutation-based approaches may be more suitable choices in this situation (Chapter 2). However, the Bonferroni correction would represent a viable and relative accurate methodology if the corrections are based on the 'effective number' of independent marker loci, rather than on the total number of variations tested. The effective number of independent loci can be estimated by different approaches based on the concept of linkage disequilibrium between markers and the observed genotypes (Nyholt, 2004; Li and Li, 2005; Gao, Starmer and Martin, 2008). Gao, Starmer and Martin (2008), for example, used a matrix of correlations between SNPs to calculate the effective number of independent tests. Their procedure is based on the extraction of the principal components (Ringnér, 2008) of the correlation matrix. Statistical corrections based on this method have shown a performance similar to that obtained from permutation-based procedures and do not require specific statistical assumptions, such as HWE (Gao, Starmer and Martin, 2008).

The PLINK software tool (Purcell *et al.*, 2007) allows the analysis of hundreds of thousands of SNPs (and CNVs) for thousands of individuals. It offers different algorithms for summarizing data, hypothesis testing and association analysis. Moreover, PLINK offers powerful solutions to correct for multiple-hypotheses testing and population stratification (Chapter 2). PLINK also allows the analysis of potential (non-random) genotyping failure. Using a typical genome-wide association study dataset consisting of 100 000 SNPs and more than 300 individuals, PLINK can load, filter and implement a complete association analysis in less than one minute (Purcell *et al.*, 2007). Statistical analysis options include HWE tests, estimation of inbreeding coefficients for each individual, and various statistical tests to detect differences between groups, such as the Chi-2, Fisher's exact and Armitage trend tests. Bonferroni and FDR methods are also offered for multiple-testing corrections (Chapter 2).

## 4.5 CNV data analysis: additional concepts, approaches and applications

Specific CNVs have been associated with common Mendelian and complex disorders, such as colour blindness, Charcot–Marie–Tooth disease, lupus, human immunodeficiency virus, Parkinson's disease, mental-retardation syndromes and Alzheimer's disease (Lupski, 2007). There is already strong evidence to indicate that many CNVs can be used to explain an important proportion of common phenotypic variation in humans, which may also be accounted for their roles in differential gene expression (Freeman *et al.*, 2006).

Different DNA chip technologies allow researchers to measure changes in copy numbers of specific parts of the genome, which opens up new possibilities to study the effect of specific deletions, insertions and duplications on disease susceptibility (Carter, 2007). As in the case of SNPs, a major challenge for bioinformatics is to improve the power and reliability of inferred associations between CNVs and specific phenotypes using different data analysis methodologies.

In a pioneering study, Redon *et al.* (2006) completed a first version of the human CNV map based on the analysis of DNA variation in 270 individuals in four populations with European, Asian and African ancestry. SNPs genotyping arrays and clone-based comparative genomic hybridization were used to scan for CNVs in these DNA samples. The resulting CNV map consisted of 1447 overlapping and adjacent variable regions, which included hundreds of genes, disease loci and other functional elements representing about 12% of the genome. Two important findings derived from this and other recent studies are that many of these CNVs are strongly inter-related through linkage disequilibrium, and that different populations exhibit major differences in copy numbers.

In a more recent study, Perry *et al.* (2008) applied a high-resolution, array-based comparative genomic hybridization platform to analyze the genomic DNA sequences of 30 individuals from the HapMap project. Their study focused on known CNV regions and showed that the size of 1020 of these regions, out of 1153, had been overestimated in previous studies (Perry *et al.*, 2008). Moreover, they found that approximately 8% of the CNV regions observed in different individuals can be characterized by a complex organization of smaller CNVs nested into larger ones. Such informational diversity and complexity of emerging findings underline the key roles that advanced applications of pattern recognition and statistical analysis will play in CNV analysis in the near future.

Advanced machine learning approaches are already contributing to the application of CNV information to novel diagnostic and prognostic models. Using CNV data acquired with array-based comparative genomic hybridization, Rapaport, Barillot and Vert (2008) proposed a supervised classification methodology based on support vector machines (Chapter 3), which was tested on different cancer datasets. Domain knowledge was incorporated into the classifiers to aid in the reduction of the feature relevance search space and the complexity of the resulting models (Chapter 3). Such domain knowledge was encoded in the form of prior hypotheses about the potential relevance or redundancy of the different genomic regions represented in the datasets. This research suggests that related methodologies can be successful in the CNV-based classification of individuals and in the identification of potential clinically-relevant chromosomal regions.

## 4.6 Key problems and challenges

As discussed in Chapters 2 and 3, the accurate definition of phenotypes, as well as the construction of case and control datasets, is a critical factor that can affect the outcomes of a biomarker discovery study. This may become even more problematic in late onset diseases, such as cardiovascular diseases, because an individual initially assigned to the control group may eventually become a case in the future. Moreover, it has been suggested that the use of invasive or cutting-edge techniques for defining phenotypes may result in a selection bias towards asymptomatic patients (Topol *et al.*, 2006). These problems in turn make association studies more difficult to reproduce as the definition of case-control inclusion criteria may vary. Emerging advances in diagnostic technologies, such as imaging, can contribute to a more accurate and meaningful definition and categorization of phenotypic classes in different diseases.

The estimation of the optimal number of samples in genome-wide association studies depends on the biomedical problem and population under investigation, as well as on other experimental factors, such as tagging selection strategy. Fortunately, there are techniques available to estimate sample size and statistical power in genome-wide association studies using representative sets of genotyped and tagged SNPs. These estimations also take advantage of information publicly available, such as that generated by the HapMap project, and information about linkage disequilibrium in the data investigated. Based on such approaches, researchers have shown that statistical power is directly proportional to increases in sample sizes, as expected. Moreover, it has been observed that tagging and the selection of tagged SNPs can influence these estimations (Klein, 2007). Also it has been suggested that statistical power may be improved by genotyping more individuals at fewer SNPs, rather than having more SNPs genotyped for fewer individuals. An example of a statistical power estimation technique (with software publicly available) was proposed by Klein (2007), who defined three main analysis steps. First, for each genotyped SNP, the best tag SNP is identified. Second, the statistical power for detecting associations for each of the SNPs is computed. And finally, the average power over all the SNPs analyzed is used to define the global statistical power of the study. An introduction to sample size estimation and statistical power is given in Chapter 2.

Machine and statistical learning techniques are becoming useful tools to support the discovery of significant genotype-phenotype associations (Ziegler, König and Thompson, 2008). For example, Malovini *et al.* (2009) used Bayesian networks to infer SNPs-phenotype associations and for classification of samples in a hypertension dataset. Their key contribution was a strategy that integrates multiple SNPs (observed in the same gene) into a single 'meta-variable'. The potential of this methodology was demonstrated by comparing its classification performance against Bayesian networks built on the relevant (original) SNPs and on haplotypes independently. The estimation of the prediction performance was done with a hold-out methodology (Chapter 3), and the mean accuracy of all the methods implemented was under 65%. Another example that illustrates the potential of machine learning is the application of decision trees and random forests (Chapter 3) to explore associations between SNPs and risk in cancers. In a case-control study, Xie *et al.* (2005) detected associations between oesophageal cancer risk and 61 SNPs, which allowed classification of samples with an overall accuracy and sensitivity above 85%. The proposed methodology also comprised missing-SNP imputation, classification performance estimation through 10-fold cross-validation and the assessment of the relative statistical significance of the SNPs.

As discussed above, the integration of biomarkers with relative weak associations with a phenotype is a key requirement in studies based on genomic variation data. One reason is that a genomic variant with a relative weak marginal effect may be clinically relevant under different genetic or environmental contexts (Altshuler, Daly and Lander, 2008). Moreover, it is important to investigate potentially relevant between-marker associations that may be responsible for disease pathogenesis or risk reduction. Such relationships between allelic markers are known as epistatic interactions (Cordell, 2002). These associations and their relation to specific phenotypes have been investigated using different statistical and machine learning approaches (Loscalzo, 2007). Examples include applications in which information derived from multiple SNPs is used as the input to classification models, such as principal component analysis (Gauderman *et al.*, 2007) and neural networks (Curtis, 2007). Mechanic *et al.* (2008) developed the Polymorphism Interaction Analysis tool (PIA), which implements different techniques for estimating the potential relevance of different combinations of SNPs in relation to specific biological pathways and disease status. This system has been applied to explore gene-gene and gene-environment interactions in cancer research (Mechanic *et al.*, 2008). PIA can be seen as a feature selection system (Chapter 3), which scores the classification power of combinations of user-defined SNPs based on the statistical analysis of genotype-phenotype tables constructed from the population sample. Such tables describe the $(9 \times 2)$ possible relations between the observed genotypes and phenotypes (Mechanic *et al.*, 2008).

The diversity of types of interactions (biomarker-biomarker, biomarker-disease) and the systems-level complexity emerging from such relationships suggest that the application of network-based approaches (Chapter 7) will play a major role in future genotype-phenotype association studies. But independently of advances in computational prediction methods, the greatest challenge will be the demonstration and interpretation of the biological relevance of the epistatic interactions identified (Cordell, 2002; Hirschhorn and Daly, 2005).

Recent progress in the development of bioinformatics resources for supporting genome-wide association studies includes graphical visualization-based approaches. An example is Goldsurfer2, which allows the detection of global and local patterns based on hierarchical representations of the data and the integration of multiple views. Golden Helix (2009) is another software tool that offers different solutions to implement genome-wide analysis of SNPs and CNV data based on graphical and interactive visualization of informational patterns, such as linkage disequilibrium. Further advances in this direction are expected in the near future, and will take advantage of open-source and extensible software models, as well as of more robust statistical analysis methods (Buckingham, 2008).

Future advances in bioinformatics and their potential contributions will depend on the availability of larger datasets, better quality of assays of genomic variation relevant to both common and rare variants, as well as more detailed and standardized definitions of phenotypes.

As discussed in Chapters 2 and 3, the assessment of the classification or predictive capability of emerging biomarker models is crucial for the development and deployment of a new generation of clinical decision-support tools based on 'omic' data. Different concerns and caveats, such as the misuse or misinterpretation of the AUC (Chapter 3), also apply to the specific context of genomic variation research (Pepe and Janes, 2008). Recent discussions orientated to genotype-phenotype association research confirm the importance of using diverse, context-dependent model evaluation measures for clinical prediction, risk assessment and disease classification (Pepe and Janes, 2008).

The reproducibility of genotype-phenotype association studies will continue to be both a crucial requirement and challenge. Despite the fact that many of the associations reported to date have not been successfully validated, and that there is no general agreement on what represents a replication study, important advances to address these challenges have been recently accomplished. For instance, the NCI-NHGRI Working Group on Replication in Association Studies (2007) has defined a checklist and specific recommendations to guide the reporting and replication of genotype-phenotype association studies. This covers different aspects: reporting of study design, data access, genotyping and quality control procedures, reporting of results, implementation of replication studies, and recommendations for reviewers to assess the relevance of new findings. Examples of specific recommendations are the need to use a second experimental platform to evaluate or validate associations, and the reporting of associations involving markers in strong linkage disequilibrium with the markers putatively associated. Furthermore, the importance of disseminating negative association results and sufficient information on the implementation of the replication study has been underlined. This working group has also offered recommendations of direct relevance to bioinformaticians. This includes the detailed reporting of several aspects, such as departures from Hardy-Weinberg equilibrium, assessment of potential population stratification, data displays and documentation of methodologies. Chapter 10 will further discuss challenges and recommendations for reporting biomarker studies.

# Guest commentary on chapter 4: Integrative approaches to genotype-phenotype association discovery

**Ana Dopazo**
*Genomics Unit, CNIC, E-28029, Madrid, Spain*

This commentary focuses on the utility of integrative genomics approaches for refining our understanding of genotype-phenotype associations, specifically through the use of global gene expression analyses as a complementary approach to DNA variation studies for the identification of new clinical biomarkers.

As discussed in Chapter 4, over the last few years, new approaches to genetic mapping have yielded great progress towards the mapping of loci involved in susceptibility to common human diseases. However, although improved genetic mapping and the amount of DNA variance explained will continue to grow in the coming years, many of the genes and mutations underlying these findings still remain to be identified and the genotype-phenotype correlation in most diseases, both monogenic and common complex diseases, has yet to be well characterized.

Thus, although comprehensive genome-wide DNA analyses are now possible and additional work is underway, this only represents a first step forward towards a better biological understanding and clinical applications. To pursue this challenging task there is room for useful tools such as genome-wide expression studies. Indeed, although

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

association methods can define regions in the genome containing the genetic variants underlying pathological processes and other phenotypes, on their own they provide little insight into the functional variants and/or mechanisms underlying the phenotype. However, transcriptional profiling technologies potentially allow not only the interpretation of functional effects of DNA variants, but also the description of functionally important variants (Stranger and Dermitzakis, 2006). In this regard, transcription level is a quantitative phenotype that is directly linked to DNA variation (genotype). Although DNA polymorphisms located outside of coding regions may have no known or evident functional effects, they may directly modify gene transcript abundance through cis-regulatory regions or by altering transcript stability or splicing.

Furthermore, several studies have described the genetic basis of transcriptional variation and have shown not only that gene expression is a heritable trait (Dixon *et al.*, 2007), but also that degrees of differentiation can be found at the level of gender (Lawniczak *et al.*, 2008) and within and between populations (Zhang *et al.*, 2008).

Furthermore, current robustness of genome-wide methodologies for studying gene expression has enabled transcriptome studies on an unprecedented scale and mRNA abundance can be measured consistently in human tissues and cell lines, as further discussed in Chapter 5 of this book. Although the potential of transcriptome analysis by ultra-high-throughput sequencing is currently being explored, microarray technology is today the most widely used methodology for transcriptome analysis and several thousands of papers describing data from expression microarrays are published each year. Today the technology is considered a robust one, and several publications have demonstrated that good reproducibility can be achieved across laboratories and platforms. It has also been demonstrated that fold change results from microarray experiments correlate well with results from assays such as quantitative reverse transcription PCR. Furthermore, and although microarray expression data analysis continues to be a challenging step, basic standards have been established in this area over the last few years and points of consensus have emerged about the general procedures that warrant use and elaboration.

Within medicine, microarray-based gene expression profiling has been used successfully for cancer diagnostics. The clinical utility of array-based gene profiles has been evidenced by studies showing that cancer gene-expression signatures may affect clinical decision-making in, for instance, breast cancer and lymphoma management (van't Veer *et al.*, 2002). In this regard, it is worth mentioning that since the initial applications of expression microarray technology in the field of cancer, more than a decade ago, the US Food and Drug Adminitration (FDA) issued in February 2007 its first approval of a multigene prognostic test, the MammaPrint, based on gene expression patterns. MammaPrint, developed by the Amsterdam-based company Agendia, uses a 70-gene signature to classify women with breast cancer into 'low' and 'high' risk of metastasis (Couzin, 2007). Integrative genomic studies have already shown their potential in the oncology field. A recent review by Witte (2009) discusses how GWA studies and expression array results can together refine our understanding of prostate cancer genetics, and have implications in the screening and treatment of this disease.

The relative maturity of transcript-profiling techniques has also led to their integration into the field of cardiovascular biomarker discovery. Aiming to explore the systems biology of cardiovascular diseases by generating useful molecular genetic 'signatures' of different types of diseases, gene expression analyses have been performed on myocardial tissue to identify specific patterns in cardiac hypertrophy, myocardial infarction and different forms of heart failure (Gerszten and Wang, 2008). However, and in contrast to the relatively easy availability of tumour samples, the validation and application of transcriptional approaches to the identification of new cardiac biomarkers in humans is, clearly, limited by the availability of the relevant tissues: the heart and blood vessels. As an alternative, the use of blood as a tissue surrogate has proved successful in an increasing number of gene expression studies (Kang *et al.*, 2006) and has received increasing interest from other biomedical areas (Chapter 4).

Blood transcriptomic approaches to study cardiovascular diseases have been performed using RNA from whole blood, from different cell-types of fractionated blood and from immortalized blood cells. As shown in Chapter 5 of this book, there are an increasing number of examples in the literature of the use of this non-invasive source of clinical material and of how the peripheral blood transcriptome can dynamically reflect system-wide biology. Overall, it represents a convenient, rigorous and high-throughput method of gene expression profiling. It is quite significant that, despite the delay in the application of emerging genomic technologies in the cardiovascular arena in comparison with cancer research the FDA approved AlloMap for marketing in August 2008. AlloMap is a non-invasive test based on molecular expression techniques for heart transplant patients. This test, which is developed by the Californian company XDx, measures white blood cell gene expression values of 20 different genes related to the immune system. This molecular signature helps clinicians to monitor (post-surgery) heart transplant patients for potential organ rejection, which is a significant risk for patient survival. AlloMap is the third multigene-expression test cleared by the FDA after Agendia's MammaPrint, approved in February 2007, and 'Tissue of Origin', a microarray-based test from Pathwork Diagnostics approved in July 2008. Tissue of Origin determines the type of cancer cells present in a malignant tumour.

Advances in human genome annotations, marked improvements in gene expression technologies, with the inclusion, for instance, of genome-wide level detection of alternative transcripts and decreased cost of high-throughput experiments, is accelerating our knowledge of human genome-wide transcription in the context of genomic organization. Overall it is clear that disease susceptibility is mediated by changes in gene expression, and that the study of these changes can help us to identify pathways in which genetic variation contributes not only to common diseases, but also to rare monogenic disorders. Global transcriptome studies of DNA mutation carriers can help decipher the biological basis of disease (Oprea *et al.*, 2006).

Beyond integrative genomic studies, the identification of new biomarkers will depend not only on the complementary power of genetics and transcriptional profiling, but also on proteomics and metabolomics. This will be a long journey and an arduous transition from the research environment to routine clinical practice. However, there is no doubt that the current trend of new technologies to systematically assess variation in genes, RNA, proteins and metabolites will impact different areas of clinical practice and will contribute to personalized medicine. Despite ongoing advances, the major challenge is how to integrate different types of data into a comprehensive 'systems' view of disease in humans.

## References


Couzin, J. (2007) Amid debate, gene-based cancer test approved. *Science*, **315**, 924.

Dixon, A.L., Liang, L., Moffatt, M.F. *et al.* (2007) A genome-wide association study of global gene expression. *Nat Genet*, **39**, 1202–1207.

Gerszten, R.E. and Wang, T.J. (2008) The search for new cardiovascular biomarkers. *Nature*, **451**, 949–952.

Kang, J.G., Patino, W.D., Matoba, S. and Hwang, P.M. (2006) Genomic analysis of circulating cells: a window into atherosclerosis. *Trends in Cardiovascular Medicine*, **16**, 163–168.

Lawniczak, M.K., Holloway, A.K., Begun, D.J. and Jones, C.D. (2008) Genomic analysis of the relationship between gene expression variation and DNA polymorphism in Drosophila simulans. *Genome Biology*, **9**, R125.

Oprea, G.E., Kröber, S., McWhorter, M.L. *et al.* (2008) Plastin 3 is a protective modifier of autosomal recessive spinal muscular atrophy. *Science*, **320**, 524–527.

Stranger, B.E. and Dermitzakis, E.T. (2006) From DNA to RNA to disease and back: the 'central dogma' of regulatory disease variation. *Human Genome*, **2**, 383–390.

van't Veer, L.J., Dai, H., van de Vijver, M.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

Witte, J.S. (2009) Prostate cancer genomics: towards a new understanding. *Nature Reviews. Genetics*, **10**, 77–82.

Zhang, W., Duan, S., Kistner, E.O. *et al.* (2008) Evaluation of genetic variation contributing to differences in gene expression between populations. *American Journal of Human Genetics*, **82**, 631–640.

# 5 Biomarkers and gene expression data analysis

This chapter will introduce gene expression data analysis in the context of biomarker discovery. Fundamental statistical concepts and problems for disease classification and prediction model design will be reviewed. This will be followed by a discussion of recent advances and applications in different medical application domains. The content will be guided by the following topics: (a) biomedical findings and clinical applications, (b) statistical and data mining methodologies applied, (c) strengths and limitations.

## 5.1 Introduction

Changes in gene expression can be measured by different types of techniques ranging from smaller to large-scale approaches, and differing in terms of their reliability and genome coverage: Northern blotting, real-time polymerase chain reaction (RT-PCR), serial analysis of gene expression (SAGE), multiplex PCR and different types of DNA microarrays. These tools allow the detection of differentially expressed genes, up- or down-regulated genes in relation to specific clinical conditions or functional pathways. These studies may also be expanded by follow-up or validation studies using additional gene expression data measured with alternative experimental platforms, or by the implementation of other 'omic' approaches, such as proteomic approaches. The large-scale acquisition of gene expression data has allowed the design of different biomarker models for diagnostic and prognostic applications in cancer, cardiovascular diseases and other pathologies.

Most of the early biomarker discovery studies in this area consisted of the comparison of two classes of samples to detect statistically important differences in gene expression, and which aimed to support the prediction of disease emergence or progression (Quackenbush, 2006). van't Veer *et al.* (2002) reported a pioneering study on the application of biomarkers based on gene expression profiling for supporting the prediction of clinical outcomes. The main goal of this study was to discover a set of biomarkers to classify (same stage) cancer patients according to their response to therapy. The result was a set of 70 genes, whose expression profiles were powerful enough to infer the clinical outcome in young individuals.

Significantly differentially expressed genes have been traditionally used as the starting point of different biomarker discovery and validation investigations. In principle, such differential patterns can be directly used as potential biomarkers if their association with the phenotype investigated offers enough discriminative power. Their applicability may be enhanced when the gene expression patterns are strongly correlated with the expression of proteins, especially those that can be measured in the blood or other fluids. Traditional gene expression analysis for cancer and cardiovascular biomarker discovery has comprised the profiling of *in vitro* or *in vivo* tissue from tumours, explanted hearts or biopsies. More recently, the application of whole-blood or plasma-based gene expression profiling has been proposed as a novel alternative to biomarker discovery.

In the areas of cardiovascular diseases, advances in gene expression analysis have allowed the identification of a variety of potential biomarkers, such as those useful to distinguish between ischaemic and non-ischemic heart failure, and between hypertrophic and dilated cardiomyopathies (Kittleson and Hare, 2005; Rajan *et al.*, 2006). Gene expression profiling has also allowed the identification of putative biomarkers of atherosclerosis, atherosclerotic lesions, plaque rupture, vascular stress and vascular remodelling. More recently, the gene expression analysis of peripheral blood cells has become a promising approach to identifying powerful biomarkers and less-invasive diagnostic techniques, as well as to assessing treatment effects and dissect key molecular mechanisms involved in the development of coronary heart disease (Patino *et al.*, 2005; Chittenden *et al.*, 2006). The systematic analysis of gene expression patterns identified in tumour samples has also allowed researchers to propose several novel biomarkers associated with different tumour types and responses to treatments (van't Veer and Bernards, 2008). Moreover, gene-expression biomarkers have enabled the identification of sub-classes of cancers and prognostic signatures in breast and lung cancers. The combination of microarrays and genomic variation analysis through sequencing of diverse genes (Chapter 4) has enabled a more detailed characterization of individual cancers (Sawyers, 2008).

Examples of commercially available biomarker systems based on gene expression data are MammaPrint, Oncotype DX and the H/I test (van't Veer and Bernards, 2008). MammaPrint, by Agendia, is a 70-gene signature for breast cancer prognosis; Oncotype, by Genomic Health, consists of 16 gene-expression biomarkers for predicting the recurrence of breast cancer patients treated with an aromatase inhibitor; and the H/I test, by AviaraDx, is a 2-gene signature that is used to estimate the risk of recurrence and response to therapy of breast cancer patients.

## 5.2 Fundamental analytical steps in gene expression profiling

Advances in experimental technologies for measuring the abundance of messenger RNA have triggered important changes in the way data are stored, organized, retrieved, analyzed and shared in the laboratory and clinical environments. Most of the gene expression data analysis techniques in biomarker discovery research are based on widely-investigated methodologies and tools developed in the areas of statistical analysis, computational intelligence and data mining over the past 40 years. However, domain- and user-specific requirements, constraints and goals have more recently motivated the development of new methodologies, tools and resources specifically tailored to gene expression data analysis.

The majority of the early studies reported potential biomarkers extracted from clustering analysis, such as different versions of hierarchical clustering (Quackenbush, 2006). However, currently the main role of unsupervised classification analysis (Azuaje, 2003;Wang, Zheng and Azuaje, 2008) may be better defined as an approach to exploratory data analysis, which is typically followed by supervised classification modelling (Azuaje and Dopazo, 2005). Because of this, as well as their general applicability to different diagnostic and prognostic applications, this chapter will focus on problems, techniques and advances based on supervised learning models (Chapter 3).

Figure 5.1 summarizes a typical biomarker discovery process based on gene expression data analysis. Genome-wide gene expression measurements in tissue or fluid samples with known phenotypes (e.g. metastasis vs. non-metastasis, pathology vs. control) can be obtained using different gene expression profiling techniques and

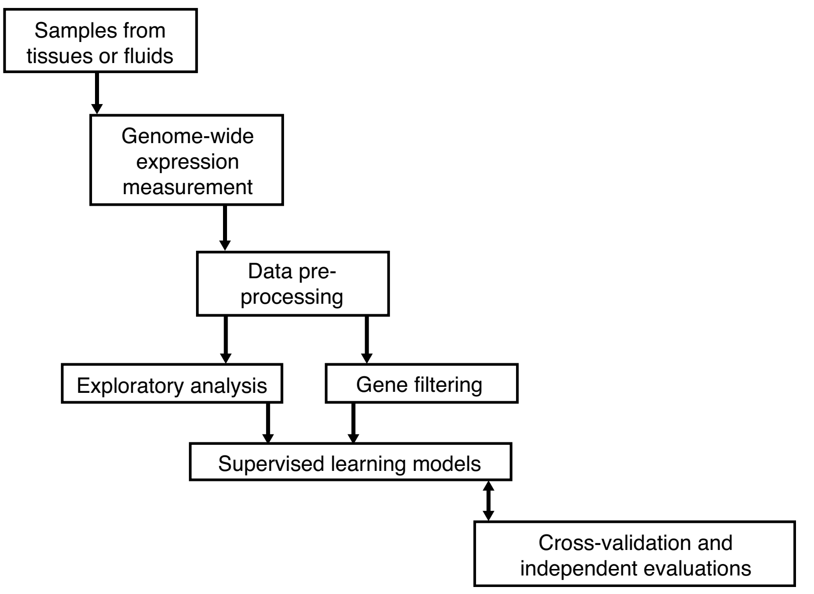


**Figure 5.1** Major phases and procedures for biomarker discovery based on the analysis of gene expression profiles

platforms. Data pre-processing includes different steps of normalization and filtering. Normalization procedures are applied to transform the data into a format that is compatible or comparable between different samples or assays, as well as to level potential differences caused by experimental factors, such as labelling and hybridization. Data filtering procedures are also needed to exclude genes or samples according to user-defined criteria, such as the exclusion of genes with very low variance across sample groups or of those genes with little correlation with the class to be predicted (Quackenbush, 2006).

The resulting dataset can be represented by a data matrix, whose rows and columns can represent the genes and samples under analysis respectively. Under this data representation scheme, a row would represent a vector of expression values associated with a particular gene across the different samples (e.g. tumour samples), and a column would represent a vector with the expression values of the genes from a specific sample or patient. Standard data visualization procedures include colour-coded representations of absolute or relative expression levels. In such procedures a colour spectrum can range, for instance, from green through black to red. This colour scale can be used to reflect the range of gene expression values observed in a dataset: from the lowest negative (down-regulation) to the maximum positive (up-regulation) values.

After performing standard normalization and pre-processing procedures, a typical analysis task is to determine which genes can individually be used to distinguish between two groups of individuals. This can be done by applying, for instance, several hypothesis testing approaches using different expression or fold change thresholds and several criteria for making multiple-hypotheses testing corrections (Chapter 2). This phase can also be supported by exploratory data analysis based on unsupervised classification (clustering) and visualization. Different filtering approaches (Chapter 3) can also be applied to further remove uninformative, highly noisy or redundant genes for subsequent analyses. This is commonly followed by the implementation of a variety of supervised classification techniques, including those that can be used to perform ‘wrapper’ and ‘embedded’ feature selection (Chapter 3). Despite different efforts to compare many of the supervised classification techniques commonly applied (Li, Zhang and Ogihara, 2004a; Pirooznia *et al.*, 2008; Statnikov *et al.*, 2005; Statnikov, Wang and Aliferis, 2008), there is no single approach that can be considered as the ‘best solution’ to the great diversity of biomarker research applications based on gene expression profiling. For additional guidance on design factors and selection criteria the reader may refer to Chapter 3.

The classification or predictive capability of the resulting models is estimated by cross-validation, followed by different independent validation phases (Chapter 3). In practice, this biomarker discovery process relies on the analysis of gene expression data together with multiple sources of information, including literature, as well as curated functional annotations (e.g. Gene Ontology terms) and disease-related pathways stored in different external public and commercial databases. Figure 5.2 depicts such an integrative view of gene expression analysis for biomarker discovery, in which prior knowledge is used to: (a) describe potentially relevant genes and processes, and (b) to estimate the potential biological and clinical relevance of the outcomes. The best known example of this approach is the statistical detection of Gene Ontology (GO) terms, mainly biological processes, significantly represented in a set of genes differentially expressed across case-control groups (Al-Shahrour *et al.*, 2006). The typical approach

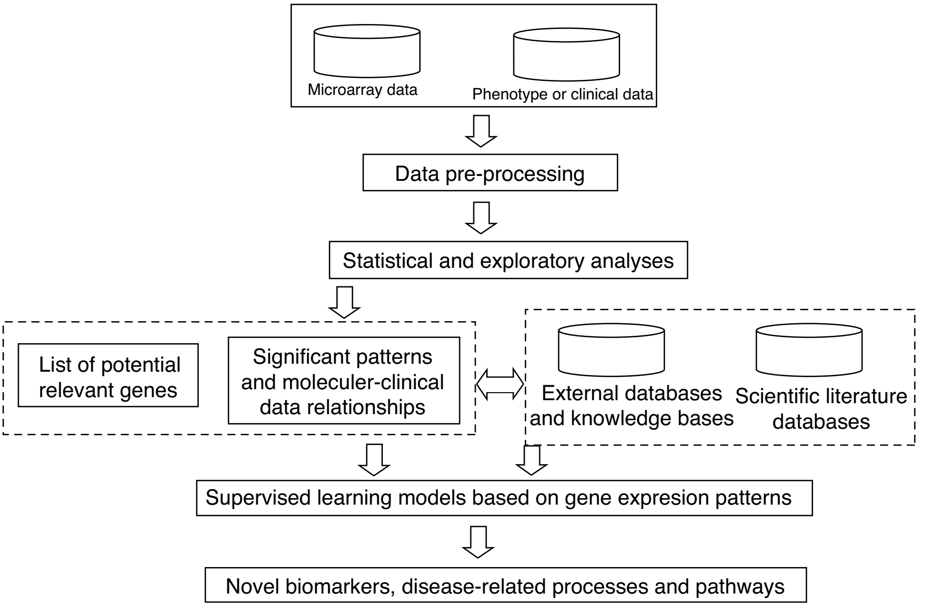


**Figure 5.2** A typical example of the integration of gene expression data analysis with prior knowledge to support the identification, characterization and evaluation of potential biomarkers

identifies sets of genes overrepresented in a specific functional category (Goeman and Bühlmann, 2007) using statistical tests, such as those based on $2 \times 2$ contingency tables and different methods for multiple-hypotheses testing corrections (Chapter 2).

Data pre-processing steps are essential for dealing with the following problems: digital re-formatting, data encoding, missing data, filtering and data transformation. The latter includes different procedures for re-scaling, normalizing and standardizing the data. These procedures are fundamental to assure data integrity and quality prior to their integrative analysis and modelling (Azuaje and Dopazo, 2005). Data exploration and statistical analyses are in practice implemented through several iterations. This allows researchers to focus their attention on smaller sets of potentially novel and interesting data patterns (e.g. groups of samples or gene sets). This in turn is useful to confirm initial hypothesis about the relevance of the features available and to guide future experimental and computational analysis. This can be achieved by the application of different data reduction, mapping and visualization-based techniques (Azuaje and Dopazo, 2005). As explained in Chapter 2, the implementation of statistical analysis tasks is fundamental to characterize major differences between groups of patients, such as presence vs. absence of a disease, or other clinically-relevant sub-classes. This can require the application and interpretation of parametric and non-parametric hypothesis-testing tasks for different types of data (e.g. numerical, nominal).

The resulting groups of genes and associations derived from initial exploratory phases are analyzed by the combination of information extracted from publicly or commercially available data sets, annotated functional databases and published papers. These tasks aim to: (a) estimate the potential relevance of the identified genes and relationships; (b) to discover other significant genes and relationships (e.g. gene-gene or gene-disease) not found in previous data-driven analysis steps. Some of the external databases and tools that can support these tasks are: human gene annotation databases (including those annotated to the GO), metabolic pathways databases (e.g. KEGG), gene-disease association extractors from public databases (e.g. Endeavour), and functional catalogues. These tools are further discussed in Chapters 8 and 9.

These information sources can be explored, integrated and exploited through different open-source software platforms, such as Cytoscape and GEPAS (Chapter 9). Large-scale data mining and information extraction from published papers, such as those indexed in Medline, can also be implemented by combining the application of human expert-driven verification and different advanced text mining tools, such as iHOP (Hoffmann and Valencia, 2005), PreBIND (Donaldson *et al.*, 2003) and Chilibot (Chen and Sharp, 2004).

The combination of the resulting data- and knowledge-driven findings, patterns or predictions provide a selected catalogue of genes, pathways and (gene-gene and gene-disease) relationships relevant to the phenotype classes investigated. The analysis and integration of this diversity of gene expression-based analysis outcomes can also be seen as part of a systematic, integrative data mining process that requires the incorporation of diverse and functionally-related information resources (Chapter 8). Such iterative, incremental discovery processes also aims to computationally generate and experimentally validate new hypotheses about the biological roles of new biomarkers or their value as potential therapeutic targets.

## 5.3 Examples of advances and applications

Microarrays may be used to measure the level of expression of thousands of genes in parallel in different samples types and specific experimental conditions. In cancer research these technologies have been widely investigated for supporting molecular classification of different types of tumours and responses to treatment (Azuaje and Dopazo, 2005; Montero-Conde *et al.*, 2008). Within the area of cardiovascular diseases, their application for aiding in the understanding of heart failure (HF) in non-ischemic, inherited (dilated and hypertrophic) cardiomyopathies has received relatively more attention (Barth *et al.*, 2006; Rajan *et al.*, 2006; Wittchen *et al.*, 2006) in comparison to ischaemic (post-myocardial infarction) heart failure (Stanton *et al.*, 2000). Furthermore, research in the HF area has mainly focused on the study of end-stage HF (Tan *et al.*, 2002; Sanoudou *et al.*, 2005; Benjamin and Schneider, 2005; Frankel *et al.*, 2006).

The application of microarray technologies has helped to provide new insights into known disease-related processes or to determine novel molecular mechanisms relevant to the development of HF (Tan *et al.*, 2002; Seo, Ginsburg and Goldschmidt-Clermont, 2006; Wittchen *et al.*, 2006; Heidecker *et al.*, 2008). The involvement of cytoskeleton-related genes, as well as genes relevant to energy production, contractility, cell cycle control, immunology and apoptosis, has also been associated with different stages of HF (Grzeskowiak *et al.*, 2003; Liew, 2005). For example, modifications in gene expression patterns connected to energy metabolism have been consistently observed across dilated (DCM), hypertrophic (HCM) and ischaemic cardiomyopathies (ICM) (Kittleson *et al.*, 2004; Sanoudou *et al.*, 2005; Sharma *et al.*, 2005). In addition to these molecular pathways and genes, recent studies have shown significant roles of protein translational, matri-cellular and immunological mechanisms in end-stage HF (Sharma *et al.*, 2005).

Other studies have confirmed the down-regulation of immune response genes in DCM as prominent indicators of HF (Nian *et al.*, 2004; Barth *et al.*, 2006).

Microarrays have also been applied to distinguish gene expression profiles of ICM and non-ICM (Kittleson *et al.*, 2004). It has been shown that gene expression profiling can accurately predict cardiomyopathy etiology, which also supports the application of microarray technologies for HF prognosis. The comparison of multiple studies indicates that, although different etiologies underlying HF may be characterized by common end-stage patterns of gene expression, it is also possible to identify unique patterns directly correlated with the etiology or the severity of the disease (Nanni *et al.*, 2006; Heidecker *et al.*, 2008). These technologies have also been successful in finding statistically-detectable differential changes of gene expression in coronary artery disease, such as those observed between atherosclerotic vs. non-atherosclerotic samples (Archacki *et al.*, 2003; Patino *et al.*, 2005).

Major alterations of gene expression patterns in response to myocardial infarction have been identified in rat models of MI (Stanton *et al.*, 2000). More than 200 genes have been shown to be significantly differentially expressed in response to MI in the left ventricle and intra-ventricular septum, in comparison to samples obtained from normal samples. More compact and robust sets of genes have been proposed for the classification of different HF etiologies (ICM and DCM) and normal hearts (Barth *et al.*, 2006). Such signatures include the deregulation of genes encoding brain natriuretic peptide (BNP), BNP-related, sarcomeric structural, cell cycle, proliferation and apoptosis proteins in cardiomyopathy.

The biomarkers presented above were identified mainly by applying statistical hypothesis testing procedures (Chapter 2), including multiple-hypotheses testing, based on traditional parametric methods, such as the t-test and ANOVA, applied to relatively small sample sets (<100). This can be seen as a defining feature of a 'first generation' of gene expression-based biomarker studies in different biomedical areas. This gave way to relatively more powerful and reliable biomarker selection and classification techniques, such as SAM (Chittenden *et al.*, 2006; Popper *et al.*, 2007), PAM (Wang *et al.*, 2007), decision trees (Huang *et al.*, 2008), random forests (Boulesteix, Porzelius and Daumer, 2008) and support vector machines (Montero-Conde *et al.*, 2008). These and related statistical and machine learning techniques were introduced in Chapters 2 and 3.

Blood cells represent a novel and promising source of molecular information, which can be assessed through microarray data analysis. The ability of blood RNA to reflect molecular and physiological states of solid tissues and organs in humans have been demonstrated (Liew, 2005; Moore *et al.*, 2005). The predictive potential and applicability of this resource is rooted in the fact that there is a continuous dynamic interaction between blood cells and the body organs. This interaction may induce subtle changes in the gene expression patterns of the blood cells, which actually mirror physiological modifications or stimuli at the tissue or organ levels.

This strongly indicates the relevance of biosignatures extracted from blood RNA as potential 'biosensors' to estimate, for example, the presence or future onset of a disease. It has been estimated that blood cells can express approximately 80% of the genes encoded in the human genome (Liew *et al.*, 2006). A similar proportion of genes expressed in different organs, including the heart, have also been detected in peripheral blood samples (Liew *et al.*, 2006). Diagnostic or prognostic applications of blood cell gene expression profiling have been evaluated in a diverse range of diseases, such as coronary heart disease (Ma and Liew, 2003), hypertension (Bull *et al.*, 2004), Kawasaki disease (Popper *et al.*, 2007), development of collateral circulation in patients with coronary heart disease (Chittenden *et al.*, 2006), different types of cancer (DePrimo *et al.*, 2003; Montero-Conde *et al.*, 2008), lupus (Bennett *et al.*, 2003), hepatitis C virus infection (Huang *et al.*, 2008) and neuronal injuries (Tang *et al.*, 2003).

Eady *et al.* (2005), for instance, demonstrated that transcript levels for the majority of genes in peripheral blood cells are consistent within samples from the same individual, and that variation between healthy individuals may be statistically detectable or 'significant'. Also their study showed that between-individual differences in expression profiles can be explained by sex, age and body mass index. Another important conclusion was that gene expression profiles obtained from an individual can be comparatively stable over time. Eady *et al.* (2005) arrived at these conclusions after analyzing gene expression profiles with hypothesis testing procedures (*t*-test, ANOVA) and multiple-hypotheses testing corrections based on the Bonferroni and FDR methods (Chapter 2).

The potential clinical significance of approaches to detecting blood-derived bio-signatures is enhanced when one takes into account that blood sample extraction is relatively non-invasive and inexpensive in comparison with many traditional procedures, such as biopsies. Thus, blood sample gene expression profiling represents a feasible substitute for solid tissue samples, such as those directly obtained from tumour or heart tissue, which can significantly accelerate translational biomedical research and the implementation of more advanced clinical decision-support systems.

Early contributions in this area have gone beyond the application of the typical *t*-test and hierarchical clustering approaches, and incorporated more rigorous methodologies to implement multiple-hypotheses testing and supervised classification. One such an example was the demonstration that peripheral blood mononuclear cells from patients during acute ischaemic stroke show differentially perturbed gene expression profiles (Moore *et al.*, 2005). Following the application of t-tests and corrections for multiple-hypotheses testing based on Bonferroni and FDR methods (Chapter 2), the PAM algorithm (Chapter 3) detected a set of 22 genes as potential strong predictors of stroke. The classification model construction process was implemented on samples from stroke and healthy individuals (20 samples in each group). The classification model was subsequently validated on an independent cohort of similar sample size, and resulted in sensitivity and specificity values around 80%.

## 5.4 Examples of the roles of advanced data mining and computational intelligence

Some of the examples introduced above are based on the application classification and prediction models originating from the area of data mining and computational intelligence, as introduced in Chapter 3. Different schemes that integrate these models with traditional statistical analysis, such as the multi-stage, serial application of filters prior to machine learning classification, have been reported in the literature. These advances aim to address typical design obstacles, such as the curse of dimensionality and the availability of noisy or incomplete datasets.

An example of the application of a hybrid 'learning' methodology was based on the combination of 'partial least squares' data dimensionality reduction (PLS) and random forests (Boulesteix, Porzelius and Daumer, 2008). This approach also integrated gene expression and clinical information, such as traditional prognostic factors, to demonstrate the added prediction value of gene expression information in cancer prognostic applications. In this example, random forests were built on reduced datasets obtained from the microarray and clinical datasets independently. The authors demonstrated how microarray-based biomarkers can add significant predictive value in comparison to or in combination with traditional clinical biomarkers. Another multi-stage filtering and classification methodology was reported by Huang *et al.* (2008) to predict the treatment response of patients infected with the hepatitis C virus. The outcomes of several statistical hypothesis-testing techniques, such as the $t$ test, Wilcoxon test and SAM (Chapter 2), were combined to detect lists of differentially expressed genes between 'good' and 'poor' response groups. The resulting genes were used to construct a decision tree-based classifier (Chapter 3).

Gene expression profiles measured with microarrays in peripheral blood cells have also been linked to the occurrence of thoracic aortic aneurysm (TAA) based on the combination of traditional statistical analysis and machine learning techniques. For example, Wang *et al.* (2007) proposed a gene expression signature that can accurately detect individuals at risk of developing this pathology. A 41-gene classification model achieved overall classification accuracy above 75%. An independent validation of the model on gene expression data acquired with real-time PCR reported similar results.

Wang *et al.*'s methodology (2007) can be summarized as follows. The model training phase (Chapter 3) enabled feature selection and classification model construction based on multiple-hypotheses testing and the PAM algorithm (Chapters 2 and 3). The dataset consisted of 61 samples: 36 TAA patients and 25 controls. The data acquisition and standard pre-processing phase resulted in more than 16 000 genes, which were subsequently analyzed for ranking purposes based on a bootstrap sampling method and the t-test. In this gene ranking (multiple-hypotheses testing) process, multiple datasets were randomly generated from the (training) data by sampling with replacement (Chapter 3) and the t-test was applied to each of these partitions. For each data partition, Wang *et al.* (2007) selected and recorded the top 500 genes on the basis of their corresponding $P$ values. An overall gene ranking procedure can be conducted by looking, for example, at the frequency of the ranked genes across all the partitions or at the average rank of each gene.

Based on this method, Wang *et al.* (2007) obtained a list of around 100 genes that were differentially expressed between TAA and control classes. These genes then represented the input to the PAM technique that selected a set of optimal features for classification, using 10-fold cross-validation on the complete training dataset (Chapter 3). The resulting PAM-based 41-gene classifier reported overall (10-fold cross-validation) classification accuracy, sensitivity and specificity above 75%. This classifier was tested on an independent dataset consisting of 33 samples (22 TAA and 11 controls), and reported similar classification performance results.

## 5.5 Key limitations, common pitfalls and challenges

These emerging and future advances strongly motivate and require the application of advanced data analysis techniques and tools. New research directions are also needed to augment the quality and range of applications of microarray technologies. This would not only demand a deeper understanding of the underlying mechanisms of health and disease, but also the identification of novel diagnostic and prognostic biomarkers. However, the current state of the area also highlights fundamental limitations and obstacles to the full realization of the benefits promised by these technologies and advances. Key experimental challenges are the complexity and cost of extracting samples from heterogeneous tissue, as well as the possible inability to represent alternate splicing, post-transcriptional modifications and reflect the activity of specific cellular localizations. The level of noise and variation of data in real-world applications can also be influenced by errors and inconsistencies in sample and assay handling, as well as differences in intra- and inter-laboratory experimental conditions and assay processing protocols (Simon, 2006).

From a computational research point of view, inconsistent and relatively insufficient statistical analyses, may represent major sources of false positive predictions (Ginsburg, Seo and Frazier, 2006; Azuaje and Dopazo, 2005; Jafari and Azuaje, 2006). The heterogeneity of sample sources and experimental protocols may also strongly deteriorate or confound the observed responses or behaviours (Ginsburg, Seo and Frazier, 2006). Furthermore, there is a need for applying more accurate and reliable methods for measuring or defining phenotypic classes (Chapter 10).

Important limitations created by the biological and statistical nature of gene expression data should be considered when designing models and interpreting results. Major challenges and obstacles include the fact that gene expression levels may significantly vary not only between disease states, but also between the samples defining these groups. In addition, important variations between samples can be actually influenced by experimental factors, such as the selection of 'controls' for data normalization in different experimental microarray analysis platforms (Tanriverdi and Freedman, 2008).

Potentially spurious differences in expression profiles may also be explained by the origin of the biological samples analyzed, rather than molecular effects directly related to a disease or phenotype. For instance, it has been reported that for the same biomedical study important differences may be observed between datasets obtained from whole-blood and leucocyte samples (Tanriverdi and Freedman, 2008). It has been shown that, when these two gene expression data acquisition methods are compared, seemingly 'significant' differential expression patterns may be determined by the biological source of the RNA, and not by the actual processes or disease conditions under comparison (Feezor *et al.*, 2004).

The power of biomarkers for the prediction of responses to treatments, such as cancer therapies, can be limited by the fact that some treatments may produce very subtle changes in gene expression and because specific responses may be driven by multiple, subtle variation patterns at the DNA level (Chapter 4). This motivates the investigation of novel computational advances to support the implementation of classification and prediction models based on heterogeneous data inputs, such as the combination of gene expression and genomic variation data. Also, it has been suggested that hundreds of samples are required to discover potentially useful gene expression signatures for drug response prediction (van't Veer and Bernards, 2008). In cancer research, the inclusion of metastatic cancer patients undergoing multiple treatments may represent another major problem because of the difficulty in establishing specific treatment-response associations. These challenges require context-specific decisions in connection to the selection of patients, therapies and clinical settings (van't Veer and Bernards, 2008).

The problems associated with multiple-hypotheses testing, the curse of dimensionality and the relative lack of data will continue receiving special attention as critical design factors (Chapter 2). This will be required despite the variety of 'standard' approaches and software tools available, and the increasing 'awareness' of such problems in the life and medical sciences research communities. In some cases, perceived limitations may be easily addressed by improving reporting practices or providing more detailed information about experiments and biomarker models (Chapter 10). However, in some cases there are still reasons to be concerned about the soundness of computational and statistical methodologies applied (Simon, 2006). Researchers should continue improving practices to reduce predictive bias and model over-fitting through the correct application of cross-validation approaches. The latter should include a clear separation of training and testing (and evaluation) phases with regard to the datasets selected. This and other problems related to selection bias were discussed in Chapter 3. The main application principle is that the data used to build a classification model, which may also include a 'wrapper-based' feature selection phase, should not be used to test the model or estimate its predictive performance. An example of the correct application of this fundamental practice is provided by Asgharzadeh *et al.* (2006), who applied nested cross-validation (Chapter 3) to build, optimize and evaluate a classification model for patients with metastatic neuroblastoma.

Chapters 2 and 3 reviewed available options for estimating and reporting model accuracy or prediction performance, which are suitable to gene expression-based classification techniques. If the desired outcome of a biomarker study is the development of new disease classification or clinical outcome prediction systems, then researchers should move beyond the idea of simply listing 'significance' statistics. This requires placing emphasis on fundamental properties relating to the discrimination, classification or prediction capability showed by the computational models built with such biomarkers. Appropriate cross-validation and independent evaluation of models on the basis of key factors, such as sensitivity and specificity (Chapter 2), should be seen as the crucial 'tests' to be passed by a new diagnostic or prognostic model. Thus, researchers should avoid over-emphasizing the potential 'significance' of *P* values or regression coefficients associated with the inputs of a statistical model, that is individual biomarkers (Simon, 2006).

The relative lack of gene expression data and the underlying biological complexity of diagnostic and prognostic models, together with the increasing availability of annotated information in the form of functional pathways and networks, are triggering the development of new biomarker selection and classification methods based on the integration of gene expression and network information. Important advances, to be reviewed in more detail in Chapters 7 and 8, have been based on the idea of searching for signalling pathways or protein complexes showing high differential expression activity in relation to specific phenotypes. A typical approach consists of focusing on those pathways with 'high differential expression scores', such as those based on t-statistics, between case and control samples (Lee *et al.*, 2008). Thus, the genes participating in such 'perturbed' pathways can be selected as the inputs to subsequent analyses or classification models based on different machine learning techniques. These approaches may not only outperform computational models based on gene expression data only, but also offer alternative insights into the molecular mechanisms underlying the pathogenesis and development of diseases or responses to therapies.

# Guest commentary on chapter 5: Advances in biomarker discovery with gene expression data

**Haiying Wang, Huiru Zheng**
*Computer Science Research Institute, School of Computing and Mathematics, University of Ulster, Newtownabbey, Co.Antrim, BT37 0QB, UK*

With the ability to measure simultaneously the expression levels of thousands of genes in a single experiment, global gene expression profiling technologies such as microarrays and serial analysis of gene expression (SAGE) offer significant advantages in the search for new biomarkers. However, the massive amounts of genome-wide expression data generated pose a great challenge for data mining and analysis. It has been shown that traditional statistical and classification techniques are not sufficient to address some fundamental issues in the search of novel and meaningful biomarkers. For example, one common practice is to apply statistical tests to score genes on the basis of their association with specific clinical outcomes and then to select the top-ranked genes as biomarker candidates, which may result in the identification of a set of highly correlated biomarkers. Gerszten and Wang (2008) argued that, in order

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

to achieve a significant improvement in predictive performance, new orthogonal biomarkers associated with new disease pathways are needed. Unsupervised clustering techniques and recent advances in network-based analysis offer great benefits in this endeavour.

## Unsupervised clustering approaches

Clustering is the process of partitioning a set of data items into clusters, such that similar items are grouped into the same cluster, whereas items in separate clusters are more dissimilar. Clustering techniques have attracted a great deal of attention in gene expression analysis because they can detect previously unknown classes in large high-dimensional data. As shown by D'haeseleer (2005) and others, data clustering is often one of the first steps in a typical gene expression analysis.

The recognized significance of unsupervised clustering has triggered many efforts to design accurate *ad hoc* clustering algorithms. In the past decade, numerous clustering algorithms have been proposed, such as: hierarchical clustering, *k*-means and self-organizing maps (SOM). However, due to the heterogeneity and complexity of gene expression data, none of the currently available clustering algorithms has consistently outperformed the others. Thus, researchers typically apply different clustering models and compare outcomes in order to generate more meaningful and reliable results. To ease this burden, a number of interactive, integrated clustering and visualization platforms have been published. Recent effort includes the *AMIC@* Web service (Geraci, Pellegrini and Renda, 2008), which provides a uniform and highly interactive interface to several clustering algorithms for gene-expression data.

It is worth noting that most of the current available clustering packages mainly focus on the analysis of gene expression data obtained from microarray experiments. Clustering analysis of SAGE data, for example, has received less attention. There are two important differences between SAGE and microarray data. Unlike microarrays, which are restricted to the analysis of previously characterized genes, SAGE allows for the detailed examination of all the transcripts present within a cell without full knowledge of gene sequences. Also, the generation of SAGE data is governed by a different statistical model. Unlike microarray data, it has been shown that the number of a specific SAGE tag observed in a specific SAGE library approximately follows a Poisson distribution (Wang, Zheng and Azuaje, 2008). Thus, the application of currently available clustering techniques with traditional distance measures, such as the Pearson correlation and Euclidean distance, to analyze SAGE data may not be appropriate. This has motivated several efforts towards the design of novel clustering algorithms specially tailored to SAGE data analysis.

Examples of new clustering algorithms for SAGE data are the PoissonHC and PoissonS techniques (Wang, Zheng and Azuaje, 2008). The basic philosophy behind these efforts is to incorporate Poisson statistics-based similarity measures into the learning process of currently available clustering algorithms, such as Hierarchical clustering and SOM. These techniques consider an important property of SAGE data: the expression levels of highly expressed tags tend to be more accurate and reliable than those of the weakly expressed tags. A review of current advances in clustering approaches to SAGE data analysis can be found in Wang, Zheng and Azuaje (2008).

In the development of clustering approaches to biomarker discovery, another fundamental issue is to evaluate the quality of clustering results to ensure that the biomarkers identified are at least statistically reliable. Examples of techniques for clustering evaluation include the development of various cluster validity indices and class-representation statistical tests, which mainly rely on indicators derived from the data under study, such as the hypergeometric distribution-based test.

While these data-driven evaluation methods have been implemented in gene expression analysis with varying success, it is evident that they are not sufficient to determine whether the clustering outcomes may be biologically meaningful. Recent years have seen a growing trend towards the incorporation of prior biological knowledge, such as functional data and biochemical pathway maps, to assess the quality of the outcomes derived from clustering analysis of gene expression data. A comprehensive review of these applications was provided by Khatri and Drăghici (2005).

## Module-based approaches

The outcomes of traditional methods of gene expression analysis are lists of potentially relevant genes, whose relationships are inferred from the data under consideration. Though successful in many areas, such a methodology exhibits several limitations that hinder its performance. For example, because the analyses are carried out at the gene level, they are sensitive to the inherent noise that exists both in the sample population and in different data acquisition stages. Segal *et al.* (2005) argued that simply listing genes associated with certain types of diseases is far from the identification of the biological processes in which these genes are involved and the causal mechanisms that might give rise to diseases.

In an attempt to extract high level and more interpretable expression patterns associated with disease phenotypes, several studies have utilized ‘gene modules’ as the basic building blocks to study disease mechanisms at the molecular level. Instead of examining the expression profile of each gene in isolation, the idea is to analyze the joint behaviour of a set of genes and to organize them into ‘higher-order modules’, in which sets of genes act in concert to perform a specific function (Segal *et al.*, 2005). It has been demonstrated that such a module-level analysis can offer a better understanding of the molecular basis of human disease development and progression (Segal *et al.*, 2005).

Apart from examining transcriptional changes in genes or functional modules for the identification of disease biomarkers, recent contributions include the integration of gene expression data with genomic and proteomic data. One example is the combination of gene expression data with human protein interaction networks. Taylor *et al.* (2009) recently examined the ‘dynamic structure’ of human protein interaction networks to determine whether changes in the organization of the interactome can be used to predict patient outcomes. In the context of cardiovascular diseases, Camargo and Azuaje (2008) integrated gene expression analysis with a protein-protein interaction network to investigate potential biomarkers of dilated cardiomyopathy (DCM). The main outcome of their study was a set of integrated, potentially novel DCM signature genes, which may be used as reliable disease biomarkers. The reader is referred to Chapters 7 and 8 which discuss these approaches in more detail.

## Final remarks

The relative lack of robust statistical analysis and the heterogeneity of data sample sources and experimental protocols represent important challenges to discovering clinically-relevant and reproducible biomarkers. This motivates the design of new computational methods that go beyond the analysis of top-ranked genes (Ein-Dor *et al.*, 2006). This also means that new biomarker candidates may originate from lists of genes that are not necessarily differentially expressed, and that their behaviour should be analyzed at different network levels.

Module-based gene expression analysis has already provided important insights into the biological mechanisms underlying various diseases. However, there is a need to incorporate temporal and spatial information into such network-based analyses. It has been shown that the full potential of module-based analysis may not be realized without studying the dynamic organization of biological networks (Taylor *et al.*, 2009). This will benefit from the development of new integrative bioinformatic platforms that go beyond the 'single-marker' analysis paradigm (Azuaje, Devaux and Wagner, 2009).

## References

Azuaje, F., Devaux, Y. and Wagner, D. (2009) Computational biology for cardiovascular biomarker discovery. *Briefings in Bioinformatics*, **10** (4), 367–377.

Camargo, A. and Azuaje, F. (2008) Identification of dilated cardiomyopathy signature genes through gene expression and network data integration. *Genomics*, **92** (6), 404–413.

D'haeseleer, P. (2005) How does gene expression clustering work? *Nature Biotechnology*, **23**, 1499–1501.

Ein-Dor, L., Kela, I., Getz, G. *et al.* (2006) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics*, **21** (2), 171–178.

Geraci, F., Pellegrini, M. and Renda, M.E. (2008) AMIC@: All MIcroarray Clusterings @ once. *Nucleic Acids Research*, **36** (Web Server issue), W315–W319.

Gerszten, R.E. and Wang, T.J. (2008) The search for new cardiovascular biomarkers. *Nature*, **451**, 949–952.

Khatri, P. and Drăghici, S. (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics*, **21**, 3587–3595.

Segal, E., Friedman, N., Kaminski, N. *et al.* (2005) From signatures to models: understanding cancer using microarrays. *Nature Genetics*, **37** (Suppl), S38–S45.

Taylor, I.W., Linding, R., Warde-Farley, D. *et al.* (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, **27** (2), 199–204.

Wang, H., Zheng, H. and Azuaje, F. (2008) Clustering-based approaches to SAGE data mining. *BioData Mining*, **1**, 5.

# 6 Proteomics and metabolomics for biomarker discovery: an introduction to spectral data analysis

This chapter will begin with an introduction to proteomics and metabolomics: Fundamental definitions, problems and key applications, with an emphasis on data obtained from spectral analysis of clinically-relevant samples. This will be followed by a discussion on the characteristics of the data and information generated in these areas, and of key approaches to biomarker discovery in proteomics and metabolomics. An introduction to feature transformation and selection will be provided, which complements the content of Chapter 3. The chapter will conclude with an overview of key computational resources and a discussion of current challenges and emerging research directions. The overview of resources will be complemented by Chapter 9.

## 6.1 Introduction

Proteomics and metabolomics have become promising technologies for the discovery of biomarkers in different complex multi-factorial diseases, such as cancers (Abate-Shen and Shen, 2009) and cardiovascular diseases (Sabatine *et al.*, 2005). These areas refer to the analysis of the clinically-relevant catalogues of proteins and metabolites. These approaches may represent powerful complementary views of the molecular state of the

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

cell at a particular time. One of the major challenges is the diversity of cell types contributing to the human proteome and metabolome (e.g. measured in the plasma) and the low concentration levels of many of the proteins suggested as potential disease biomarkers.

The wide range of biomedical applications of proteomics and metabolomics means that there is a continuous growth of the volumes and complexity of these data, which will require advanced algorithms, methodologies and software for data analysis, management and visualization (Radulovic *et al.*, 2004). This in turn will need the development of new approaches to information representation and modelling, and advanced tools for data reusability and interoperability (Cannataro, 2008).

The aim of proteomics is to identify and quantify all proteins expressed in a cell at a given time or under specific biological conditions. Metabolomics is the study of the products of metabolism, such as sugars, lipids and amino acids. The metabolome is defined as the complete set of metabolites found in the human body. Such metabolites represent molecules that are smaller than proteins, RNA and DNA molecules. Small molecules typically refer to molecules with mass smaller than 1500 Daltons (Da) (Wishart, 2007). A related area is ‘metabonomics’, which places emphasis on the systemic changes of complex systems through time (Nicholson and Lindon, 2008). But commonly the terms metabolomics and metabonomics are used interchangeably.

Proteomics and metabolomics share many computational problems, requirements and challenges with other key ‘omic’ areas, such as genomics and transcriptomics. Biomarker discovery based on any of these data types depends on the availability of software and data analysis techniques for selecting relevant features, classification and management in the presence of the curse of dimensionality and multiple data formats (Chapter 3). Moreover, progress in biomarker discovery is driven by statistical and machine learning techniques, knowledge management and integrative software infrastructures.

This chapter focuses on major data mining tasks required for disease biomarker discovery based on spectral data, which can originate from proteomic or metabolomic experiments. Special emphasis will be given to computational problems, concepts and ‘generic’ approaches to analying spectral data, which can be common, or easily adapted, to both areas. Key limitations and requirements of techniques and applications are also discussed.

## 6.2 Proteomics and biomarker discovery

Proteomic databases, biomarkers and applications have been reported to classify patients on the basis of different phenotypes or medical conditions. Biomarker discovery using proteomics has traditionally relied on the identification of differentially expressed proteins using control (e.g. healthy, good treatment response) and case (e.g. disease conditions) samples secreted in serum, plasma or solid tissue (Anderson, 2005; Peacock *et al.*, 2008). Examples of technologies for identifying or measuring protein expression are: Western blotting, 2-dimensional gel electrophoresis, antibody arrays, mass spectrometry and nuclear magnetic resonance (Section 6.4). Amongst them, the most widely applied technique is mass spectrometry (Webb-Robertson and Cannon, 2007). In practice, different combinations of experimental technologies (Section 6.4) may be required for extracting proteomic spectral data from different types of tissue and fluid samples. These experimental technologies have allowed the generation of large sets of proteomic data from human tissue, seminal fluid, urine, blood, and cell lines (Duncan and Hunsucker, 2005; Peacock *et al.*, 2008).

Over recent years proteomics research has developed two main experimental directions: The identification of peptides/proteins in samples, and the understanding of physiology and pathology based on large-scale datasets. The former task has been based on the search and retrieval of similar peptide/proteins from databases and *de novo* predictions (Webb-Robertson and Cannon, 2007). The latter includes protein profiling that aims to discover biosignatures, for example from urine or plasma samples, capable of differentiating between disease and control groups (Duncan and Hunsucker, 2005).

The main phases involved in a typical biomarker discovery process based on mass spectrometry are: (a) sample preparation, separation and labelling; (b) experiment implementation and data acquisition; (c) spectra pre-processing; (d) peptide/protein identification (database or de novo sequencing) and quantitation; and (e) pattern discovery and classification (Cannataro, 2008). The latter two tasks, (d) and (e), do not require sequential implementation: the implementation of these tasks actually will depend on the study goals and resources. For instance, pattern discovery and classification are suitable for large-scale analysis of potentially novel biomarkers described by sample fingerprints or signatures. This phase is challenged by the need to implement sound and robust data dimensionality reduction, transformation or selection (see below), in which the number of features representing each sample is much larger than the total number of samples available (Chapter 3).

Data pre-processing includes signal filtering, baseline subtraction, normalization, noise reduction, peak extraction, dimensionality reduction and transformation, and data exploration (Veltri, 2008). The first five tasks are commonly implemented using commercial software embedded with the data acquisition equipment. The other pre-processing tasks may be implemented by combining a variety of commercial and public software tools.

Database search methods have become the traditional method for peptide/protein identification, and an increasing number of advanced databases and search algorithms are becoming available to augment the accuracy of identification tasks, together with greater reduction of false positive rates. The database search approach to peptide identification requires a database of known organism-specific peptides, which is used to match an unknown spectrum in the experimental samples under investigation. Thus, the 'most similar' spectra in the database is used to infer the peptide to be associated with the experimental spectrum based on different 'matching' or 'scoring' metrics. When different matching spectra are found in the database, a 'spectrum model' can be estimated to be compared with the experimental spectrum (Webb-Robertson and Cannon, 2007). Different types of information are used to estimate between-spectra similarity, such as peak intensities and correlations between the spectra. The inference of spectra models can also be obtained from average probability values, individual probability values associated with each ion in a spectrum, or peak occurrence patterns detected by statistical learning approaches (Webb-Robertson and Cannon, 2007). These techniques together with advances in data compression are improving the capacity of matching unknown peptides to sets of studied sequences stored in different databases. For supporting the identification of proteins in proteomic database search methods

(Section 6.3), there are several commercial and non-commercial systems that mainly query publicly-available databases (Veltri, 2008). However, these and other applications are becoming available as modules of integrated solutions that are being developed to implement the different phases of the proteomic data management workflow (Section 6.8).

As expected, the association of an experimental spectrum with a peptide using the database search approach depends on the availability of known peptides accurately represented in the database. Moreover, its performance will be deteriorated in the presence of novel proteins, mutations, complex mixtures of proteins and experimental (sequencing) errors. *De novo* peptide identification aims to address these difficulties (Veltri, 2008). This approach is based on the assumption that a spectrum is encoded by a set of ions that can be used to estimate, in principle, the mass of the peptides. This, together with information on the distance between the ion peaks, can be used to estimate the peptide composition. In practice, these methods generate information on partial sequences only due to low mass accuracy and incomplete fragmentation of peptides in the experimental sample (Webb-Robertson and Cannon, 2007). Methods based on statistical learning, graph theory and optimization algorithms have been used for *de novo* peptide identification (Dancik *et al.*, 1999; Chen *et al.*, 2001; Heredia-Langner *et al.*, 2004; Webb-Robertson and Cannon, 2007). Graph theoretic methods represent peaks and mass differences as network nodes and edges respectively, and peptide identification is done through network path analysis (Dancik *et al.*, 1999; Chen *et al.*, 2001). Optimization methods aim to match an experimental spectrum (from the unknown protein) with an amino acid sequence by means of the maximization (or minimization) of an optimization (or fitness) function. Different fitness functions and optimizations algorithms, for example genetic algorithms (Heredia-Langner *et al.*, 2004), have been proposed. A disadvantage of *de novo* identification methods is that they may retrieve incomplete or partial sequence information. Database identification methods also tend to be more user-friendly (or understandable) and to offer more options to constrain the search space.

As discussed in Chapter 5 in the case of gene expression analysis, the circulatory system offers a great source of potential disease biomarkers because peripheral blood serum and plasma contain a great variety of relatively abundant proteins that may reflect diverse physiological (or pathological) states of different organs and body responses. This potential is enhanced due to the relatively less-invasive nature of these tests. Moreover, different studies have demonstrated that tissue-derived proteins can also be directly measured in plasma by mass spectrometry (Hanash, Pitteri and Faca, 2008). Even before the era of large-scale proteomics, plasma- and serum-derived proteins provided important diagnostic and prognostic biomarkers for different cardiovascular pathologies (Arab *et al.*, 2006), such as BNP (Chapter 1), and cancers (e.g. prostate-specific antigen). A variety of protein biomarkers relevant to ovarian, pancreatic and colon cancers can be measured in serum by today's proteomic profiling technologies (Hanash, Pitteri and Faca, 2008), and it is very likely that in the short term new biomarkers will be discovered in other medical areas, such as neurological disorders, using large-scale proteomic approaches. The combination of different experimental platforms, together with more advanced computational methods for feature extraction, transformation and selection, are making it possible to detect novel and sensitive proteins differentially expressed across control and pathological states.

## 6.3 Metabolomics and biomarker discovery

Initial estimates of the size of the human metabolome indicate that there are a few thousand endogenous metabolites, that is those synthesized by enzymes encoded in the human genome (Wishart *et al.*, 2007). Through metabolomics, researchers aim to improve their understanding of the mechanisms distinguishing health and disease, and of differential responses to treatments (e.g. side-effects, effectiveness) (Abate-Shen and Shen, 2009). This is motivated by the fact that metabolites can reflect important changes in the activity of genes and proteins, with small changes at the gene or protein levels having larger effects on the concentration of metabolites (Pearson, 2007). Two examples of the application of metabolite measurements in modern, routine clinical practice are the analysis of cholesterol and glucose levels for diagnostic purposes in heart disease and diabetes. Such an 'early' introduction in traditional health care enhances the translational research potential of post-genome-era metabolomics (Van and Veenstra, 2009).

The search for disease biomarkers would be greatly supported by the availability of information cataloguing the majority of metabolites. Despite ongoing efforts (see below) that have collected, described and stored thousands of human metabolites and their associations with different diseases in databases (Abate-Shen and Shen, 2009), this has become a task more complex than many researchers anticipated. One reason is that the number of metabolites in the human body varies according to the source of the samples (e.g. urine, blood) and the methodology used to detect the metabolites. Moreover, one has to distinguish between the great variety of metabolites produced by the human body and those generated by gut bacteria, food and drugs (Pearson, 2007). Important variations can also be affected by gender, age, the time of sample acquisition, dietary preferences (e.g. vegetarians vs. meat eaters), different environmental factors (e.g. stress and anxiety) and biogeographical origin of the patient (Van and Veenstra, 2009). Because of this, in some studies, patients are required, for example, to fast and abstain from smoking prior to sample collection (Mayr, 2008). Moreover, it has been shown that differences in the storage time of the samples may influence sample classification (Mayr, 2008).

As with genomic variation studies (Chapter 4), researchers can investigate metabolite profiles using targeted and large-scale approaches (Section 6.6). Targeted studies focus on a relatively small set of metabolites specified in advance by the researchers. Large-scale (or pattern discovery-oriented) approaches require the discovery and analysis of massive sets of 'fingerprints' represented by complex peak spectra (Sections 6.5 and 6.7).

The analysis of spectral data from metabolomic studies comprises several preprocessing, management and classification tasks, which require the application of different software tools and statistical methods. Before feature extraction and selection procedures are implemented (Section 6.7), data normalization using different algorithms is performed as part of the pre-processing phase. Data normalization can be seen as a domain- and platform-specific problem because of the existence of different sources of experimental variation or error. The two main families of normalization methods for metabolic profiles are: Traditional statistical techniques for data scaling, and normalization based on reference compounds (Sysi-Aho *et al.*, 2007). The former normalizes each sample in relation to the whole dataset and can be based, for instance, on the median of intensities or the maximum likelihood method. Normalization based on reference compounds is implemented through rules that reflect chemical properties relating a sample to a reference compound. For instance, an optimal normalization factor for each metabolite measurement can be computed based on the variability of multiple reference compounds (Sysi-Aho *et al.*, 2007).

Several commercial and public software systems are becoming available, which provide automated implementation of pre-processing, management, visualization and classification tasks (Shulaev, 2006). However, there is a lack of tools to offer comprehensive and integrated solutions covering the range of major data analysis and interpretation phases required in metabolomics research (Shulaev, 2006).

A significant proportion of published studies from cancer and cardiovascular research (Mayr, 2008) have reported the application of procedures to transform the original data, mainly through principal components analysis (Section 6.7), into a reduced set of variables to describe the samples. This is typically followed by different types of classification approaches, such as those introduced in Chapter 3. As in the case of gene expression research (Chapter 5), these tasks are enhanced by the application of tools to study the involvement of pathways and functional annotations in a specific set of metabolites (Section 6.8 and Chapter 9).

Examples of potential clinically-relevant metabolites for cancer diagnosis and prognosis include lactate, nucleosides and lipids, which are commonly detected at higher concentrations in different tumours (Van and Veenstra, 2009). Plasma metabolite biomarkers for coronary artery disease, myocardial ischemia, heart failure and lipoprotein profiling are examples of diagnostic and prognostic applications in cardiovascular research (Sabatine *et al.*, 2005; Mayr, 2008). In the case of lipoprotein profiling, concentration differences between small low-density and high-density lipoproteins may be used to detect insulin resistance or possible associations with the progression of diabetes (Mayr, 2008).

A recent investigation demonstrated the potential of metabolomic profiles obtained from tissue, urine and plasma for the identification of prostate cancer patients with risk of metastasis (Sreekumar *et al.*, 2009). Sreekumar *et al.* (2009) combined liquid-and-gas-chromatography-based mass spectrometry (Section 6.4) to profile more than 1100 metabolites in 262 samples: benign prostate, prostate cancer and metastatic disease. Amongst the potential prognostic biomarkers identified, sarcosine was shown to be strongly elevated during cancer progression leading to metastasis. More important, their research demonstrated that differential levels of sarcosine can be detected in urine. Metabolites differentially present across the clinical categories were detected with the Wilcoxon rank-sum test (also known as the Mann–Whitney test) and false discovery rates were estimated with permutation tests (Chapter 2). A variety of well-known bioinformatics tools were applied to support data visualization and interpretation: hierarchical clustering, heat maps and mapping of the differential metabolites onto the KEGG and Oncomine Concept Map databases (Chapter 9). Classifications of independent samples were implemented using the concentration values of sarcosine only. The predictive power of this biomarker was estimated by computing AUC values (Chapter 2). Sarcosine outperformed the biomarker currently used in clinical practice, prostate-specific antigen (PSA), with AUC values around 0.70 for different clinical subgroups of patients.

## 6.4 Experimental techniques for proteomics and metabolomics: an overview

There are two major technologies for proteomic and metabolomic analysis: Nuclear magnetic resonance spectroscopy (NMR) and different versions of mass spectrometry (MS), which offer complementary advantages in terms of the amounts and types of samples required, capacity to detect smaller metabolites and sensitivity. For example, tandem mass spectrometry (TMS) allows a precise determination of proteins and metabolites in a wide range of complex fluids (e.g. serum and plasma). In this technique a first MS step is implemented, peptide peaks are selected, and a second MS run is implemented on the fragments of this peptide (Veltri, 2008).

NMR can be used to detect metabolites that contain a 'NMR nucleus', that is molecules with an odd number of both protons and neutrons (Mayr, 2008). Based on the property of 'spinning motion' of the nucleus around its axis, such nuclei produce NMR signals that are detected by NMR. These signals are represented by the frequency spectra associated with the NMR nuclei found in the sample investigated. The separation of nuclei frequencies from a reference frequency is defined as the 'chemical shift'. Such a separation of frequencies allows the identification of the different molecules in the sample. The spectra are graphically represented with the frequencies plotted in decreasing order across the abscissas axis, and with the relative concentration of nuclei reflected by the peak intensities displayed on the ordinate axis.

A key advantage offered by NMR is that the metabolite extraction and analysis process allows the preservation of samples (e.g. tissues). Other important advantages offered by NMR spectroscopy are its relatively high reproducibility and capacity to discover unknown metabolites (Van and Veenstra, 2009). However, NMR techniques are more suitable to detect metabolites in high concentrations. This lack of sensitivity (i.e. less powerful for the analysis of low-abundance metabolites), in comparison to MS techniques, has been presented as one of the major limitations of NMR approaches (Shulaev, 2006; Mayr, 2008). Figure 6.1 shows an example of a typical NMR spectrum captured by a metabolomics experiment (Parsons *et al.*, 2007).

MS estimates the mass-to-charge ratio of ions found in the sample molecules, for example metabolites. The resulting spectral graphs also include peak intensities that reflect the concentration of the different ions. MS approaches can also take advantage of prior separation of the molecules through chromatography techniques: gas chromatography (GC) and liquid chromatography (LC). These separation techniques are needed to reduce the complexity of the sample and to maximize the capture of information from different compounds (Shulaev, 2006; Veltri, 2008). MS-GC is probably the most widely applied proteomic approach to biomarker discovery in cancer research (Hanash, Pitteri and Faca, 2008). To further improve coverage, sample fractionation can also be followed by independent analyses of the resulting fractions or analyses that focus on protein subgroups, for example glycosylated proteins (Hanash, Pitteri and Faca, 2008).

Other types of MS-based techniques (Veltri, 2008), which have been widely applied to proteomics research, are: MALDI-TOF MS (matrix-assisted laser desorption/ionization time of flight MS), and SELDI-TOF (surface-enhanced laser desorption/ionization time-of-flight MS). In metabolomics, MS techniques offer higher sensitivity and coverage of metabolites than NMR. Another important advantage, in the case of the GC-MS

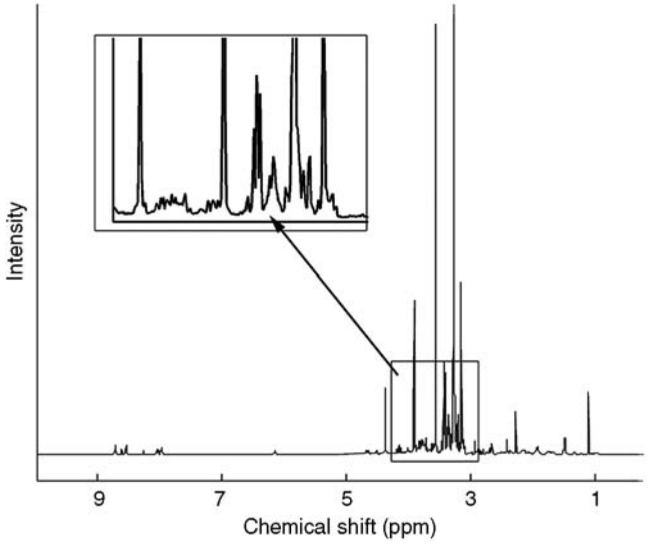


**Figure 6.1** Example of a typical NMR spectrum from a metabolomics experiment. The plot shows the NMR spectrum of a sample of mussel adductor muscle. Adapted from Parsons *et al.*, 2007, under the terms of the Creative Commons Public Domain

techniques, is the availability of a good number of databases containing mass spectral data on metabolites (Shulaev, 2006). Detailed information about these techniques can be found in (Domon and Aebersold, 2006).

## 6.5 More on the fundamentals of spectral data analysis

Spectrometric experimental techniques and subsequent data analysis allows the determination, modelling and classification of chemical compounds, including small molecules and proteins, based on their molecular weights. In the case of a mass spectrometer, it separates ions on the basis of their mass to charge ratio values. Thus, the standard output of a spectrometric experimental analysis is a sequence of value pairs: intensity vs. mass to charge ratio (m/z). Such a plot is known as the 'spectrum of the sample'. A spectrum graphically illustrates the amount and mass values of the molecules detected in the sample. Figure 6.2 shows an example of a mass spectrum obtained from patients with venous thromboembolism (VTE) (Ganesh *et al.*, 2007).

The main objectives of spectral data pre-processing are to remove experimental artefacts, reduce noise, and to ensure that the spectra originating from the different samples can be compared on a common data scale. Different pre-processing steps are required depending on the type of experimental technology, research objectives and the statistical characteristics of the spectral data obtained. For instance, typical steps involve quality control, baseline corrections, normalization, peak extraction and reduction procedures, which can be followed by different peak clustering steps for exploratory purposes (Barla *et al.*, 2008; Veltri, 2008). However, additional data dimensionality reduction may be considered as a task downstream from pre-processing because of its

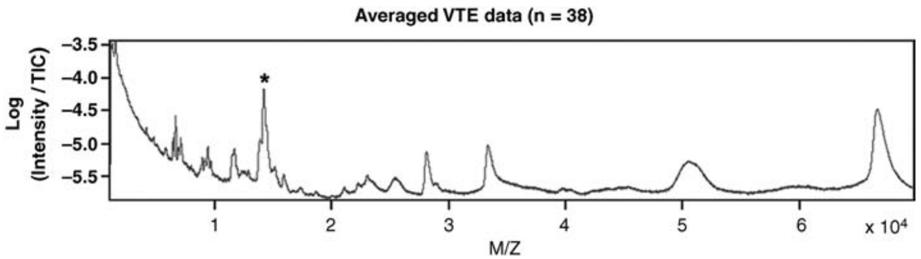


**Figure 6.2** Example of mass spectrum from a proteomics analysis. The plot shows the average spectra from 38 patients with venous thromboembolism (VTE). Adapted from Ganesh *et al.*, 2007, under the terms of the Creative Commons Public Domain

direct connection to classification and other analytical tasks (Section 6.7). Peak selection during pre-processing may be guided by different criteria: A peak value should be higher than a user-defined threshold value, it must have the highest intensity in relation to its nearest neighbouring peaks, or a peak value should be greater than a value defined by a signal-to-noise ratio (Barla *et al.*, 2008). As in the case of other 'omic' data domains, one should expect that pre-processing may significantly influence the outcomes of subsequent analytical phases, for example classification performance. In the case of proteomic data analysis, there is evidence illustrating how the choice of different data pre-processing and feature extraction techniques can affect the prediction performance of diagnostic classifiers in investigations involving synthetic and real biomedical data (Barla *et al.*, 2008).

An important pre-processing procedure in large-scale profiling investigations consists of dividing the resulting spectrum into a number of regions or 'bins', that is spectrum binning (Wishart, 2007). Binning allows 'representative' spectral values to be defined for each bin. In a MS experiment this means that, given a subset of $k$ peaks defined by pairs of spectrum (intensity, m/z) values, the spectrum peaks included in a bin will be represented by a single peak. For example, the intensity value of a representative peak can be defined as the sum of the intensities of the $k$ peaks, and the corresponding m/z value can be chosen from the set of m/z values defining the bin, for example the median value (Veltri, 2008).

Binning simplifies the computational complexity of subsequent data analysis and allows the researcher to focus on areas of specific interest to the investigation. For example, only those intensity peaks falling into a particular bin are selected for further data reduction or visualization tasks (Section 6.7). Therefore, binning can be seen as both a feature extraction procedure and a basic data dimensionality reduction method. Note that despite such a 'binning' process, the selected region of the spectrum may consist of thousands of intensity peaks. Therefore, this can be seen as a pre-selection step prior to the application of different feature reduction and transformation techniques (Chapter 3, and Section 6.7).

## 6.6 Targeted and global analyses in metabolomics

Targeted metabolomic analysis aims to identify and quantify a relatively small number of chemical compounds (metabolites) in a sample. This requires researchers to know the identities and structures of the (target) metabolites of interest (Shulaev, 2006). Targeted analyses are typically based on the analysis (i.e. matching) of its spectral data in relation to a database of reference spectra of known compounds. This is equivalent to the database search approach to protein identification introduced above. The targeted approach also assumes that the spectrum generated for a specific sample will be the spectral product of a mixture of metabolites, that is the sum of individual metabolite-specific spectra (Wishart, 2007). In the case of NMR this means that the metabolites found in the sample are expected to have unique chemical shift patterns, that is two compounds are unlikely to have the same fingerprint of peak (intensity) values and morphologies.

The methodological principles of targeted analysis can be applied to problems studied with different experimental platforms, for example different types of MS-based techniques and NMR. However, targeted analyses will be constrained by the size and quality of available databases (Section 6.8).

The main outcomes of a targeted analysis are the identity and concentration values of the metabolites found in different samples. Researchers can apply many of the statistical and machine learning techniques discussed in Chapter 3, as well as those introduced in Section 6.7. In this scenario, the inputs to the different feature reduction, selection and classification techniques may be the concentration values of the metabolites and information describing their identity or categorizations.

Global, large-scale metabolomomic analyses include studies of metabolomic 'fingerprinting' and metabolite profiling (Shulaev, 2006). It does not aim to identify and quantify target metabolites defined a priori. This approach is applied to detect profiles or bio-signatures that can be used to characterize metabolic processes in a particular sample, that is metabolic 'fingerprinting'. Different data mining techniques can be applied for data feature selection, pattern visualization and classification. Applications of this approach to biomarker discovery have used different versions of MS and NMR. The global measurement of the levels of a group of metabolites in a sample is known as metabolite profiling (Shulaev, 2006). Metabolite profiling can also be useful to support the functional characterization of genes with unknown phenotypic effects at the transcriptional regulatory levels, but whose changes or effects may be reflected on the concentration levels of metabolites (Shulaev, 2006).

## 6.7 Feature transformation, selection and classification of spectral data

The curse of dimensionality makes the reduction of data dimensionality an essential analysis phase in disease biomarker discovery (Chapter 3). In proteomic and metabolomic research a typical dimensionality reduction process aims to filter out thousands of potentially irrelevant spectral features. As further explained above, these features encode sample spectrum peaks, for example intensity values corresponding to molecular weights, associated with different peptides or compounds. Chapter 3 provided introductions to feature reduction and selection in the context of supervised classification. This section will further discuss these problems in the context of spectral data analysis, which can be applied to diverse proteomic- and metabolomic-based biomarker investigations.

Researchers apply a data transformation method for feature reduction with the aim to map data from a dimensional space of size $p$ (i.e. each sample described by $p$ features), to a transformed space of size $q$, with $q < p$. Principal component analysis (PCA) is a widely applied statistical technique for feature (dimensionality) reduction and sample visualization based on feature transformation. PCA has become a basic tool to support data analysis in biomarker studies based on proteomics and metabolomics. It aims to reduce the original set of features (intensity peaks in the case of spectral data) by means of its transformation. Such a feature transformation is done by linearly combining the original features to generate a new set of (transformed) variables. The outcome of a PCA is an optimal transformation of features that preserves the information encoded in the original data.

A disadvantage of PCA, and of other data transformation-based dimensionality reduction methods, is loss of interpretability, as the original set of features is replaced by a set of transformed variables. Also note that this and other techniques introduced in this section may not be recommended for the specific purpose of identifying molecules. They are more suitable to detect global differences and similarities between samples, which in turn may allow sample classification based on clinically-meaningful classes. In part this is because traditional feature selection techniques, for example those based on the reduction of co-linearity between features or the statistical testing of multiple hypotheses, may cause the loss of important information required for molecular identification.

The transformation obtained by a PCA aims to explain or preserve the variance observed in the original dataset. Given the original dataset, $\mathbf{X}$, composed of $p$ features and $n$ samples, the PCA finds a matrix, $\mathbf{W}$, of $p \times q$ values, which maximizes the variance of the matrix resulting from the product, $\mathbf{X} \cdot \mathbf{W}$. This is equivalent to projecting the original data onto a new set of vectors defined by the rows of $\mathbf{W}$, which are called the principal components. The principal components are actually the eigenvector values obtained from the covariance matrix, $\mathbf{C}_{\mathbf{X}} = \frac{1}{n-1} \mathbf{X} \cdot \mathbf{X}^T$. Hence, the optimization problem consists of estimating a matrix $\mathbf{W}$ that can best represent $\mathbf{X}$. This is commonly done by implementing the 'singular value decomposition' technique of the matrix $\mathbf{X}$ (Ringnér, 2008; Hilario and Kalousis, 2008). The output of a PCA indicates the relevance of the obtained principal components. This ranking is expressed in decreasing order on the basis of the capacity of the principal components to capture the variance of the original dataset. In this way the first principal component will account for the greatest percentage of the (original data) variation, followed by the second principal component, and so on.

An alternative way to explain PCA is presented as follows, using a simple hypothetical example. Suppose that our dataset consists of $n$ samples, each one described by three features (e.g. representing spectral peaks or gene expression values): $v_1$, $v_2$ and $v_3$. As explained above, PCA will create new variables, the principal components, which are linear combinations of the original data features. This means that the principal components, PC, can be represented as: $\mathrm{PC}_i = a_{i,1}v_1 + a_{i,2}v_2 + a_{i,3}v_3$, for the $i$th principal component, and with $a_1$, $a_2$ and $a_3$ representing different, constant values for each principal component. Thus, each sample can be represented by $k$ principal components, whose values are computed with this linear mathematical function, and in which $v_1$, $v_2$ and $v_3$ represent the original feature values representing a given sample.

A PCA can reduce the feature dimensionality to $n$ features without any important loss of variation information. Further reductions can support the production of visual displays or provide the inputs to relatively less complex classifiers (Ringnér, 2008), that is small number of inputs. Thus, each sample can be represented by a small number of principal components instead of thousands of values. In practice, researchers concentrate on reduced sets of principal components, for example focus on the first two or three principal components. Following this procedure, PCA allows one to produce visual displays (in 2D and 3D) of the samples based on the resulting principal components that best describe the data. This visualization can in turn suggest potentially relevant areas or patterns, that is clusters grouping clinically-meaningful sample categories, which can be subsequently analyzed with more advanced clustering or supervised classification techniques.

Examples of the crucial role of PCA in the development of (proteomic and metabolomic) diagnostic or prognostic models across different biomedical areas have been reported. A new diagnostic tool for *African trypanosomiasis* was proposed based on the PCA reduction of the dimensionality of original blood mass spectra from different patients (Papadopoulos *et al.*, 2004), which originally consisted of hundreds of features, that is spectrum intensity peaks. The resulting principal components represented diagnostic signatures, which were used as inputs to classification models capable of identifing infected patients. Different variations of PCA can be implemented to meet application-specific requirements or to exploit available information. For example, Jansen *et al.* (2004) adapted PCA to include weights reflecting the experimental error of repeated measurements of NMR. Their weighted PCA first estimated the variation of replicated measurements in a sample. Lower variation values reflect lower experimental errors. The resulting variation values were then used to define weights for each measurement, with lower variation values defining larger weights (i.e. 'more relevant' or 'reliable' measurements). This was followed by the computation of principal components. The authors showed how this weighted PCA can provide alternative views to those provided by traditional PCA (Jansen *et al.*, 2004).

Different techniques widely applied in the areas of biomedical signal processing and engineering (Sörnmo and Laguna, 2005;Clifford, Azuaje and McSharry, 2006) can also be used for proteomic and metabolomic spectral feature extraction and reduction, such as the Fourier and wavelet transforms. These techniques allow the representation of the original spectra as a linear combination of basis mathematical functions. Such analyses will also generate a set of transformed features (e.g. wavelet coefficients) that can be used for further exploratory analyses and classification model construction. In feature dimensionality reduction applications researchers can select those transformed features that meet user-defined numerical criteria. For instance, wavelet coefficients with values above a specific threshold are selected to represent a biosignature for classification purposes. In a recent investigation Alexandrov *et al.* (2009) proposed a feature selection and classification framework for diagnosing colorectal cancer based on the wavelet transformation. Using high-resolution mass spectrometry they generated proteomic data obtained from cancer and control serum samples. The original spectra data were transformed using discrete wavelets and those coefficients that showed significant discriminatory power were selected for subsequent analyses. Discriminatory power was estimated by standard parametric and distribution-free hypothesis-testing procedures (Chapter 2). The most relevant wavelet coefficients were then used to train and test, through cross-validation (Chapter 3), support vector machine classifiers. Independent evaluations of the best models were carried out on datasets generated by the MALDI-TOF technique. Remarkably, the classification accuracy, sensitivity and specificity of these evaluations were above 95%.

These feature reduction or transformation methods are unsupervised approaches because they only use the data obtained from the samples without reference to their corresponding diagnostic or prognostic classes. Supervised dimensionality reduction techniques have shown their potential in different proteomic and metabolomic applications. Examples of widely-applied supervised feature reduction techniques are Fisher's Linear Discriminat Analysis (LDA) and the Partial Least Squares (PLS) method (Hastie, Tibshirani and Friedman, 2001). LDA projects the original data onto a transformed space described by $c$-1 dimensions, where $c$ represents the number of samples classes (Hilario and Kalousis, 2008). The traditional PLS method is a regression technique (Chapter 3), which can be adapted to classification problems and implements a reduction of the original data feature dimensionality (Boulesteix and Strimmer, 2006). PLS, as PCA, comprises the identification of linear combinations of input features, which optimize the variance of the data. But unlike PCA, PLS achieves this together with the maximization of the correlation of the transformed features and the class variable. The applications of PLS and LDA, as well as their combination with other data mining approaches, have been explored in different biomarker discovery domains, including cancer research (Purohit and Rocke, 2003; Hilario and Kalousis, 2008).

The multiple-hypotheses testing procedures introduced in Chapter 2 can also be applied, as in the case of gene expression analysis (Chapter 5), to the reduction of features in proteomics and metabolomics. They include different parametric and non-parametric tests: t-test, F-ratio, Chi-2 and Wilcoxon rank test, which have supported the identification of potentially novel diagnostic and prognostic biomarkers across different biomedical domains (Hilario and Kalousis, 2008). Information-theoretic approaches, such as mutual information (Steuer *et al.*, 2002), have also been proposed for feature filtering in proteomic research (Hilario *et al.*, 2003).

Methods traditionally applied to gene expression data, such as SAM (Chapter 2) and the centroid shrinkage method (Chapter 3) can also in principle be applied to proteomic or metabolomic data. Different applications have indicated their potential to support biomarker discovery in these areas. An example, brought by the creators of the centroid shrinkage method-based PAM technique (Chapter 3), is the classification of mass spectrometry data using the 'peak probability contrasts' method (Tibshirani *et al.*, 2004). They illustrated its application using mass spectra data from an ovarian cancer investigation, which allowed them to propose a set of seven spectral peaks as potential biomarkers from a pre-processed set of 192 peaks.

Other techniques proposed for the reduction, visualization and classification of different types of spectral data include the Soft Independent Modelling of Class Analogy (SIMCA) method, the application of PCA prior to linear discriminant analysis (DA), and Partial Least Square-DA (PLS-DA) (Holmes *et al.*, 2000; Smith and Baert, 2003; Wilson *et al.*, 2005; Wishart, 2007). SIMCA is a feature reduction technique that differs from traditional PCA in the sense that automated sample clustering and classification is included, based on a training phase and cross-validation (Chapter 3). Unlike PCA, PLS-DA can also be directly used for supervised classification of samples. Although PLS-DA exploits some of the principles of PCA, it uses information about the clinical or phenotypical categorization of samples in a training dataset. This guides both the reduction of the feature dimensionality and the clustering of the training samples. The resulting model can then be applied to assign classes to unknown samples in a test dataset.

In general, feature selection techniques, including those wrapped around or embedded into different classification models, can also be applied to feature dimensionality reduction, the discovery of potential biomarkers and classification of proteomic and metabolomic data. Examples include correlation-based filtering methods, algorithms based on instance-based learning, algorithms embedded into support vector machines, and evolutionary computation techniques for finding optimal sets of features. Saeys, Inza and Larrañaga (2007) and Hilario and Kalousis (2008) provide brief introductions to several examples of the application of these supervised feature selection techniques to spectral data dimensionality reduction and classification in different biomedical areas. The reader may also refer to Chapters 2 and 3 for an introduction to different unsupervised and supervised feature selection techniques.

Different machine learning approaches, ranging from $k$-nearest neighbours and decision trees to support vector machine models (Chapter 3), have been reported for the design of novel diagnostic and prognostic systems based on proteomic and metabolomic data. Recent comparisons, for instance, using proteomic data suggested that support vector machines, with different kernels and feature selection methods, can produce the highest classification performances amongst some of the best known classification approaches (Barla *et al.*, 2008). However, more conclusive evidence of the power of the selected biomarkers and classification models through independent model evaluations is needed (Chapters 1 and 10).

## 6.8 Key software and information resources for proteomics and metabolomics

Metabolomic databases (Wishart, 2007) are being developed to support data mining of metabolites in the context of different biological processes and phenotypes across several organisms (Chapter 9). Advances are required to enhance the application of these databases to large-scale metabolite characterization studies in the context of health and diseases. This will require the integration of information on metabolites, pathways, networks, and cellular localization, as well as spectral data derived from different experimental technologies. Examples of metabolomic databases are the Human Metabolome Database (HMDB) (Wishart *et al.*, 2007), METLIN (Smith *et al.*, 2006) and Golm (Kopka *et al.*, 2005). Amongst these examples, the HMDB is probably the most comprehensive resource for supporting biomarker discovery research. It stores diverse types of chemical, spectral, clinical, genomic and phenotypic data for thousands of human (endogenous and exogenous) metabolites (Wishart, 2007; Wishart *et al.*, 2007).

'Generic' laboratory information management systems can be applied to support metabolomic and proteomic research, but more domain-specific solutions are necessary for data acquisition, tracking and management tasks. These systems are also essential to implement sample tracking, storage and processing, as well as daily laboratory management (e.g. research notebooks) in large-scale projects. Examples of laboratory information management solutions specific to metabolomics are the SetupX (Scholz and Fiehn, 2007) and Sesame (Markley *et al.*, 2007) systems. SetupX is a Web-based laboratory information management system compatible with XML information-encoding, with access to a metabolic annotation database and capable of supporting MS data acquisition and visualization tasks (Wishart, 2007). Sesame is a Java- and Web-based system, which can be applied to both MS and NMR data acquisition and management (Wishart, 2007).

Examples of the development of integrated, open-source software systems (also see Chapter 9) to meet the demands of the proteomic and metabolomic data analysis workflow include the MZmine project (Katajamaa, Miettinen and Orešič, 2006). This computing platform-independent software allows the incremental addition of new algorithms and methods tailored to different types of MS problems and applications. MZime is freely-available and can process raw data in different formats. It offers different tools for data visualization (e.g. 2D and 3D plots of spectra), peak detection and selection, and a small number of statistical analysis options (e.g. PCA). Another example of integrated software for the analysis of proteomic and metabolomic is the DOME system (Mendes, 2002; Shulaev, 2006). Research groups performing investigations in these areas can install DOME to store, visualize and manage data from different projects. DOME is a Web-based client-server application, and implements several data analysis algorithms, such as data clustering and PCA.

Apart from MZime, other tools, such as Pep3D (Li, Zhang and Ogihara, 2004a; Li *et al.*, 2004b), Msight (Palagi *et al.*, 2005) and msInspect (Bellew *et al.*, 2006), can be used for the automatic detection of MS peaks and their association with known peptides/proteins (Veltri, 2008). Other examples of tools offering interactive (2D and 3D) visualization of spectral data derived from different types of MS-based platforms are JDXview (Haider, 2008) and SpectraViewer (Cannataro *et al.*, 2007), which also support different output and graphics formats. JDXview can display spectral data originating from MS and NMR experiments and encoded in different formats (Haider, 2008).

Examples of interactive software tools for supporting the visualization, identification and annotation of metabolites are MetaFIND (Bryan *et al.*, 2008) and MetaboMiner (Xia *et al.*, 2008). Based on the outputs generated by different feature selection techniques, MetaFIND can implement post-processing tasks that aid the user in the identification of peaks and metabolites. Apart from offering different interactive displays of feature values, this system supports the assessment of correlations between features (including those initially excluded by the feature selection algorithms) and the estimation of feature relevance based on value changes across samples. MetaboMiner is a tool that focuses on the automated identification of metabolites from spectra generated by two-dimensional NMR, which is a high-resolution NMR technique. The identification of metabolites is done through the comparison of their 2D-NMR-derived spectral patterns and a database of reference patterns associated with hundreds of pure compounds. This freely-available tool provides different interactive visual displays of spectral images, database searches and compound lists.

## 6.9 Gaps and challenges in bioinformatics

In the case of directed biomarker studies in metabolomics, a greater characterization of the human metabolome is needed. Existing experimental platforms for the measurement of metabolites may also constrain bioinformatic advances. This is because there is evidence that analyses based on spectra peak intensities or areas, that is the current 'standard' analytical approach, may suffer from lack of accuracy and precision (Van and Veenstra, 2009). As in other areas of 'omic' biomarker discovery, metabolomics requires a more accurate and standard definition of phenotypes. This is essential if bioinformatic researchers are to contribute advanced computational diagnostic and prognostic models. A crucial challenge is to define 'normal' or 'reference' metabolite concentrations or signatures, which in turn is needed to define labels for implementing classification models. In particular, what is the 'normal' intra- and inter-individual range of metabolite concentrations? A problem is that it has been shown that the ranges that can be used to define 'normal' may be large and highly variable (Van and Veenstra, 2009). For instance, using urine samples obtained from women, Xu *et al.* (2005) showed that normal oestrogen metabolite concentration levels can vary 10- to 100-fold between individuals, and this is influenced by their menstrual status.

Research in bioinformatics will be required to establish quantitative links between biomarkers and specific phenotypic traits, such as the sizes of the tumours to be detected or of the areas affected by cardiac injury. Lutz *et al.* (2008) proposed a mathematical model that estimated the minimum tumour sizes required to make accurate diagnosis using two biomarkers. This type of model could be extended or adapted to include other biomarkers provided that sufficient prior physiological knowledge is available for the biomarkers. Moreover, it could be applied to support the selection and prioritisation of early diagnosis systems (Lutz *et al.*, 2008) based on their (phenotypic trait) detection sensitivity.

Advances in proteomics and metabolomics will also depend on the availability of unambiguous definitions of molecular entities. This is a problem that the genomics community have been tackling for years. Even the definition of 'metabolite' on the basis of a maximum molecular weight threshold (e.g. $<1500$ Da) may be subject to revisions as significant advances are accomplished. The evolution of community-driven data standards and vocabularies will hopefully address this obstacle. The use of XML-based data representations of proteomic data has been promoted by the proteomics community (Veltri, 2008). These formats allow spectral data to be represented in a compressed version together with metadata. In this case, metadata can describe different aspects of data acquisition and analysis: information about experimental platforms, operational details, sample descriptions, research notes and interpretations, description of data manipulation procedures, and so on.

The development of the next generation of integrated databases for proteomics and metabolomics will be greatly supported by advances in data encoding and compression. For instance, and as in other areas relating to information retrieval, clustering-based strategies can be applied to generate spectra prototypes representing groups of similar samples, which could significantly accelerate data search, retrieval and analysis. This is what Frank *et al.* (2008) recently implemented to speed-up searches in a public MS database consisting of more than 10 million spectra. Furthermore, they showed that this clustering-driven strategy, which is based on the application of a hierarchical clustering algorithm, can reduce the number of irrelevant hits to the databases. This open-source system is available online to allow users to create queries to the database using their own data (Frank *et al.*, 2008).

The need for standard nomenclatures and accurate phenotype definitions will demand from existing and emerging bio-banking projects the development and integration of formal representations of translational knowledge, such as biomedical ontologies. In the case of phenotypic representations, ontologies and information models with specific applications in bio-banking will also require to be associated with gene and protein expression patterns (Van and Veenstra, 2009).

As in other biomarker discovery areas based on 'omic' research, further advances in the representation and sharing of information for reproducing analyses, computational models and results are needed (also refer to Chapter 10). This not only refers to the publication of more detailed and accurate supplementary information sections in journals, but also to standards or protocols for digitally encoding proteomic and metabolomic experimental and analytical data. The latter particularly refers to information required for biomarker discovery and validations.

Experimental design and downstream bioinformatic analyses need to carefully consider possible sources of variability and errors, which may facilitate the detection of spurious associations or reductions of the false discovery rate. For instance, in the case of plasma biomarker analysis, different factors may contribute to the identification of statistically 'significant' relations that may have nothing to do with the disease under investigation. Examples of such factors are: differences in experimental procedures, variability in storage conditions, population stratification (e.g. by sex, age or diet), as well as physiological differences caused by inflammation, metabolic states or chronic disease (Hanash, Pitteri and Faca, 2008).

Most of the biomarker discovery investigations reported to date are based on the application of 'generic' data analysis techniques originating from statistics and machine learning, with univariate and multivariate filtering and feature transformation algorithms as the main approaches to data visualization and classification. This and the availability of larger (and more complex) datasets will motivate the design of new algorithms and methodologies tailored to more specific data- and user-driven requirements in proteomics and metabolomics. For instance, the different algorithms required in the spectral data pre-processing phase are commonly implemented independently using a combination of public and commercially-available software. The latter includes software that is embedded into the spectrometry instrument. It has been suggested that more integrated, domain-specific approaches are required to account for the variability associated with the experimental equipment and different individuals, including intra- and inter-sample variability (Ghosh *et al.*, 2008). Moreover, as in the case of gene expression data analysis, researchers are aware of the importance of implementing and analyzing experiments in replicate (Ghosh *et al.*, 2008).

Future advances will also depend on the availability of user-friendly software tools for data retrieval, analysis and integration. The latter is also motivated by the need to combine different analytical approaches and experimental technologies. There are many opportunities for advanced computational approaches to filtering, classifying and interpreting metabolomic data, including algorithms based on machine learning (see accompanying guest commentary). Most of the work in metabolomics and disease biomarker discovery reported to date has mainly relied on classical statistical techniques, such as PCA. Moreover, the majority of the applications in proteomics and metabolomics have concentrated on binary classification problems, that is the analysis of two clinically-relevant groups. However, advances to deal with multi-class prediction problems are under way (Oh *et al.*, 2008). Also there are indications that these areas are moving rapidly to meet some of the requirements and challenges posed by the era of systems biology (Van Dien and Schilling, 2006; Drake and Ping, 2007).

# Guest commentary on chapter 6: Data integration in proteomics and metabolomics for biomarker discovery

**Kenneth Bryan**
*Cancer Genetics, Royal College of Surgeons in Ireland, Dublin, Ireland*

Advances in both proteomics and metabolomics analyses over recent years have added greatly to our genome-wide view of transcription provided by the advent of microarray gene expression analysis. With proteomics, the first step in the expression of the phenotype, the transcriptome, can now be augmented with information on post-transcriptional regulation. Indeed recent advances in our understanding of post-transcriptional regulation have made measurement of this stage of expression more relevant than ever. Small, regulating RNAs known as *microRNAs* (miRNA), have now been shown to be critical regulators of translation for many proteins and to be involved in the progression of many diseases, including multiple forms of cancer (He and Hannon, 2004). These advances further enforce the assertion that transcription alone does not determine phenotype and that further information on downstream processes

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

are needed to complete the picture. Even after translation a protein may require additional post-translational modification, activation or co-factors before it can perform its designated function. Metabolomics, the global measurement of metabolites and small molecules within a cell or bio-fluid, has recently come of age thanks to advances in experimental platforms (Goodacre *et al.*, 2004). Metabolomics may be viewed as a way to measure the final effect of many proteins. For example, one end effect of the peptide hormone insulin is an increase in glucose levels in the cells of the body. The transcription, splicing, translation and post-translational modifications as well as its successful binding to its trans-membrane receptors are all required before glucose levels in the cell are affected. In a sense the ultimate success or failure of this process can only be determined by metabolite detection.

Proteomic and metabolomic views of cellular processes reveal additional links in the chain of events that leads to phenotype expression, any of which may be a key biological marker of disease or potential drug target. How best to integrate these diverse data sources to produce an improved global model that provides further insights into biological systems is still largely an open question.

## Data integration and feature selection

From a machine learning perspective, proteomics and metabolomics data provide heterogeneous *views* of the biological system. These alternative views may share some or all data objects (experiments, samples) and may harbour common or unique information about the relationships between these objects. How we integrate data from multiple views depends on the extent to which labels are available (whether it is an unsupervised, supervised or semi-supervised task) and the specific questions we want to ask of the data. Should we be satisfied that the samples in our study are fully and accurately labelled we may proceed directly to questions such as: ‘What are the most pertinent feature variables (genes, proteins, metabolites, etc.) across the various data views that discriminate between sample classes (e.g. healthy vs. disease), and how do these translate into a biological explanation?’

However, if we have unknown or ambiguous sample labels or are curious perhaps about the existence of sub-types within a sample class we may choose to re-examine the model in an unsupervised or semi-supervised manner prior to pursuing the above question. The goal is then to construct an integrated model of the biological system under study, as defined by feature information from multiple data views, which is superior and more complete than a model generated from any individual data view.

Traditionally, integration of such alternative data views was treated as a way to create a more robust consensus model of the underlying system supported by objects and object relationships shared across all views with disagreements largely disregarded (Pavlidis *et al.*, 2002). Although data integration began in this vein, technological advances enabling improvements in data quality has meant that it may now seem overzealous to discard information solely because it is not compatible across all views. The increases in quality, reproducibility and resolution of NMR spectroscopy in recent years are a perfect example of such advances in the area of metabolomics data analysis. Furthermore, the provision of complementary information is one of the benefits of examining a phenomenon from multiple perspectives via data from alternative sources or multiple platforms. For example, the discrimination of a disease sub-class with a single amino acid mutation may not be possible in proteomics data view but the downstream effects of its impaired function may become apparent within the metabolomics view. Taking the consensus approach in building an integrated global model may lead to such a sub-class and potential biological marker being overlooked.

Interestingly the increasing quality and number of data sources available across biological domains has been mirrored by recent efforts in machine learning at producing improved global models from multiple heterogeneous data sources (Berthold and Patterson, 2004). For example the PICA (parallel integration clustering) algorithm is a novel cluster analysis approach which supports the simultaneous integration of information from two or more sources with a view to building an improved global model (Greene, Bryan and Cunningham, 2008). Again such algorithms attempt to combine multiple views synergistically, producing integrated models that reveal more information about a system than those derived from individual views alone. Once we have built a satisfactory global model we may then assign class labels, examine the class discriminating feature variables and identify bio-markers and potential therapeutic targets.

In data representations, domain variables such as mRNA, amino acids or proteins may be represented by one or more measurable variables or *features* and the extraction of the set of most informative features with regard to our model is known as *feature selection*. It is interesting to speculate how standard feature selection methods might be extended to benefit from a model learned from multiple views. One might simply employ a wrapper approach to ascertain the most important features for class discrimination. Wrappers exist in many forms but the essential premise is the same, to evaluate the relevance of features by assessing their impact on the model accuracy. Co-inertia analysis (CIA) is an interesting approach that can currently be applied to identify class discriminating features across two data views (Culhane, Perrière and Higgins, 2003). CIA works by finding dimension reduction representations of the two datasets, using PCA for example, that are maximally similar. The parallel assessment of features across MS and NMR analysis (Walsh *et al.*, 2007) and proteomics and gene expression (Fagan, Culhane and Higgins, 2007) data have been performed using CIA. As opposed to building an improved integrated model, this form of data integration focuses on simultaneous feature selection and elucidation of a biological explanation.

Once a set of class discriminating features has been established, identifying the specific bio-marker (genes, proteins, metabolites, etc.) that these features represent can be far from straight forward, especially if an investigation is high-throughput and exploratory in design. Apart from the fact that the nature of some features may be unknown (novel protein or spectral peak), in some cases features may not map directly to domain variables. In NMR spectroscopy for example a single metabolite may be represented by multiple spectral peaks. In such a case an investigator may be faced with the prospect of extracting and identifying one or more metabolite signals from a large set of retrieved features, many of which may be collinear. Post-feature selection analysis applications such as MetaFIND (Bryan, Brennan and Cunningham, 2008) (Walsh *et al.*, 2007) may aid the investigator in bridging the gap between class discriminating features and biological explanation. Once a promising set of peaks has been identified by mass spectrometry or NMR it may be referenced against standard annotated signals for metabolites contained within online databases. The Human Metabolome Database (HMDB), for example, allows a user to query an unknown metabolite for annotations using the spectral peaks coordinates (Wishart *et al.*, 2007).

Successful integration of data from various experimental sources to construct a more accurate global model is critical to furthering our understanding of the underlying biological processes and the identification of significant features. This integrative approach should better promote the discovery of the fundamental causes in disease models, as opposed to detection of ancillary or downstream effects. Biomarkers derived from such models may provide more promising drug targets across various levels of biological systems.

The era of high-throughput analysis is gradually progressing into one of multi-view learning and data integration in which investigators are striving to make full use of multiple data views derived from diverse sources and alternative experimental platforms. Recent parallel developments in machine learning and biological domains will certainly aid this development. However, cross domain collaborations, which involve sharing of experimental data, knowledge and software applications, need to be continually cultivated if the area of data integration is to progress. Encouragement from publishers for the online provision of complete datasets from biological researchers and practical usable applications from bioinformaticians would certainly aid advances. The barriers to data integration and its rewards are gradually being removed; the data, skills and collaboration infrastructures are all available, with an added dash of communal will this area has a bright future.

## References

Berthold, M.R. and Patterson, D.E. (2004) Towards learning in parallel universes. Proc. of 2004 IEEE International Conference on Fuzzy Systems. Budapest, Hungary, 25–29 July 2004, pp. 67–71.

Bryan, K., Brennan, L. and Cunningham, P. (2008) MetaFIND: a feature analysis tool for metabolomics data. *BMC Bioinformatics*, **9**, 470.

Culhane, A.C., Perrière, G. and Higgins, D.G. (2003) Cross-platform comparison and visualisation of gene expression data using co-inertia analysis. *BMC Bioinformatics*, **4**, 59.

Greene, D., Bryan, K. and Cunningham, P. (2008) Parallel integration of heterogeneous genome-wide data sources. Proc. of BIBE 2008, Athens, 8–10 October, pp. 1–7.

Goodacre, R., Vaidyanathan, S., Dunn, W.B. *et al.* (2004) Metabolomics by numbers: acquiring and understanding global metabolite data. *Trends in Biotechnology*, **22**, 245–252.

Fagan, A., Culhane, A.C. and Higgins, D.G. (2007) A multivariate analysis approach to the integration of proteomic and gene expression data. *Proteomics*, **7**, 2162–2171.

He, L. and Hannon, G.J. (2004) MicroRNAs: small RNAs with a big role in gene regulation. *Nature Reviews Genetics*, **5**, 522–531.

Pavlidis, P., Weston, J., Cai, J. and Noble, W.S. (2002) Learning gene functional classifications from multiple data types. *Journal of Computational Biology*, **9**, 401–411.

Walsh, M.C., Brennan, L., Pujos-Guillot, E. *et al.* (2007) Influence of acute phytochemical intake on human urinary metabolomic profiles. *American Journal of Clinical Nutrition*, **86**, 1687–1693.

Wishart, D.S., Tzur, D., Knox, C. *et al.* (2007) HMDB: the Human Metabolome Database. *Nucleic Acids Research*, **35**, D521–D526.

# 7 Disease biomarkers and biological interaction networks

This chapter begins with an introduction to biological networks in the context of health, disease and biomarker discovery, as well as major analysis assumptions and methodological principles. Basic statistical concepts used to analyze the structure of networks and to discover biomedical relevant knowledge are defined. This is followed by an overview of the main approaches to representing and inferring biological networks. An introduction to key network-based approaches to biomarker discovery using different types of 'omic' information is presented. The last part of the chapter includes a more detailed discussion of representative examples of methods and applications, and of current limitations and challenges in this area. The next chapter will cover some of the approaches and examples discussed here in greater detail.

## 7.1 Network-centric views of disease biomarker discovery

The availability of increasing amounts of diverse 'omic' datasets together with the need to discover complex, clinically-relevant and more subtle associations between genes and disease have motivated the application of network-based biomarker discovery methodologies. In this approach a network typically consists of a set of nodes and edges, which represent the biological system components and interactions between the components respectively. Examples of network nodes are genes, proteins, drugs and diseases. The edges may encode different types of physical interactions (e.g. protein-protein, protein-DNA interactions) and gene regulatory associations (e.g. co-expression relationships). An edge is also sometimes used to describe a known or putative link between a

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

gene and a disease, or between a gene and a drug. Thus, in general, network edges reflect a functional similarity or relation between genes or proteins, which are particularly relevant to the disease or biological processes under investigation.

Network-based approaches to biomarker discovery have been proposed in several medical areas ranging from different types of cancers (Chuang *et al.*, 2007; Xu *et al.*, 2008), diabetes (Keller *et al.*, 2008), cardiovascular diseases (Camargo and Azuaje, 2007), asthma (Lu *et al.*, 2007), infectious diseases (Suthram, Sittler and Ideker, 2005), immunity (Raza *et al.*, 2008) and ageing (Chen *et al.*, 2008a). The typical products of such investigations are: descriptions of the properties of networks in a specific phenotype, identification of pathways or processes significantly perturbed or deregulated in disease states, list of biomarkers, and the extraction of sub-networks with potential predictive applications with respect to specific phenotypes.

In a typical biomarker discovery approach using gene expression data, genes are ranked on the basis of the discriminative capacity of a gene (or genes) in relation to different classes, such disease vs. health or case vs. control. Different classification and prediction models can be implemented based on the expression levels of these markers. Despite the advances demonstrated by the use of expression-based biomarkers, this prediction and classification approach deals with important obstacles, such as the heterogeneity of tissue samples and expression and genetic variability across patients and sub-populations (Lee *et al.*, 2008). Information extracted from biological networks, such as signalling or protein-protein interaction networks, can be used to classify samples based on the 'activity' or 'behaviour' of these networks, rather than only using the expression levels of lists of genes shown to be differentially expressed across phenotypes or patient groups.

The combination of different types of 'omic' data has been proposed to understand complex functional relationships and pathological responses in systems implicated in cardiovascular diseases (Bennett, Romanoski and Lusis, 2007; Camargo and Azuaje, 2007). For instance, Gargalovic *et al.* (2006) identified modules of genes relevant to inflammation based on differential responses of human endothelial cells to oxidized lipids. These modules represented highly interconnected genes, which were used to suggest potential new functional roles for different genes in important processes, such as unfolded protein responses (Gargalovic *et al.*, 2006; Bennett, Romanoski and Lusis, 2007).

Keller *et al.* (2008) proposed a gene expression network model to study diabetes induced by obesity. Based on the analysis of gene co-expression networks related to obesity, strain and age, they identified modules of genes linked to the emergence of diabetes. One such module was composed of genes implicated in cell cycle regulation, which may be used to predict diabetes. These modules were detected based on the analysis of the correlations of differentially expressed genes in different tissues. Significant associations between the co-expression network modules and specific biological processes were established through the statistical assessment of Gene Ontology terms found in the modules. The islet cell cycle module consisted of 217 genes and exhibited expression patterns associated with age. A principal component analysis (Chapter 3) of the expression data from this module showed an obesity-dependent increase of gene expression. The identification of network modules was implemented through a method originally proposed by Zhang and Horvath (2005) for the analysis of weighted gene co-expression networks.

Network-based approaches to disease knowledge discovery also facilitate the explicit representation of relationships involving environmental determinants and disease-modifying genes and processes. This type of approach may also be useful to visualize and analyze complex relationships in diseases characterized by multiple genotypes underlying a common phenotype, as well as disorders in which a common genotype can influence different phenotypes (Loscalzo, Kohane and Barabasi, 2007). Network-based approaches also show potential to aid researchers in distinguishing between genes and processes that represent drivers or initiators of systemic perturbations from those that simply reflect downstream generic responses or effects. This is especially important in biomarker discovery research, as researchers are interested in identifying early indicators of systemic perturbation or disorder. This in turn opens possibilities for finding biomarkers with potential disease causative roles and that may be further investigated as potential therapeutic targets.

Figure 7.1 depicts a hypothetical example of the investigation of disease and biomarker discovery based on biological networks and other related sources of information. Thick arrows are used to indicate the transfer of information or outputs generated by each analysis task, and are numbered to show the order of a typical sequence of tasks. In a first phase, different types of resources can be analyzed to extract and infer networks

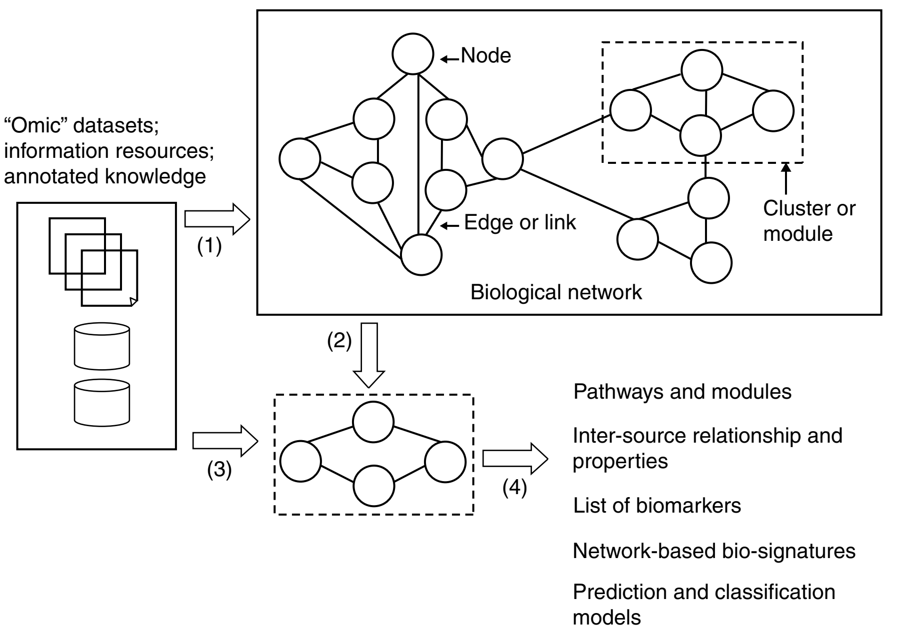


**Figure 7.1** Overview of a typical approach to investigating disease and discovering biomarkers using network-based information. Thick arrows represent transfer of information or outputs, and are numbered to indicate the order of a typical sequence of tasks: (1) Different types of 'omic' data sources, literature databases, ontology-based information repositories and curated annotations of gene-gene, gene-protein or gene-disease interactions are analyzed to extract and assemble networks of biological interactions relevant to a particular phenotype. (2) Relevant sub-networks can be extracted from the network on the basis of their potential predictive value. (3) Sub-networks may be further analysed based on their relationships with different information sources. This analysis may generate different types of outputs and testable predictions depending on the goals of the study and available resources

of biological interactions relevant to a particular phenotype or clinical response. Examples of information resources are: different 'omic' datasets, literature databases, ontology-based information repositories and curated annotations of gene-gene, gene-protein or gene-disease interactions. Basic statistical properties (e.g. clusters of nodes) that describe the structure of the network can be used to guide the search for potentially relevant biomarkers, processes or pathways. In a second phase, potentially relevant sub-networks may be extracted from the network on the basis of their possible predictive value, for example enrichment of functional processes or co-expression coordination patterns. This can be followed by another analysis phase in which the sub-networks are further dissected and interrogated based on their relationships with different information sources, such as gene expression profiles obtained from case-control groups. This discovery framework may generate different types of outputs and testable predictions depending on the goals of the study and available resources, for example list of potential biomarkers or network-based signatures for distinguishing between phenotypes.

## 7.2 Basic concepts in network analysis

A great variety of statistical concepts and indicators can be used to summarize structural properties of a node, groups of nodes or a full network. Amongst them: the degree, diameter, the clustering coefficient, the shortest path length, the characteristic path length and betweeness, are widely used to infer biologically relevant properties (Table 7.1).

The degree refers to the number of connections or interactions of a node. The clustering coefficient of a node is the proportion of possible connections between the neighbours of the node that are actually observed for a given node, that is a measure of the connection density around a node. The shortest path length of a pair of nodes in the network is the shortest distance that separates the two nodes. The diameter of a network is the length or distance of the longest of all shortest paths between a pair of nodes in the network. The characteristic path length of a network is the average value of all shortest path lengths between all nodes in the network. The betweenness of a node quantifies the number of non-redundant shortest paths passing through the node. For more detailed

**Table 7.1** Basic statistical descriptors of network structure

| Statistic | Definition |
|---|---|
| Degree | Number of connections or interactions of a node |
| Clustering coefficient | The proportion of possible connections between the neighbours of a node that are actually observed for a give node |
| Shortest path length | The shortest distance (length) separating any pair of nodes |
| Characteristic path length | The average value of all shortest path lengths between all nodes in the network |
| Diameter | The length (i.e. number of nodes or interactions) of the longest of all shortest paths between a pair of nodes in the network |
| Betweenness | The number of non-redundant shortest paths passing through a node |

mathematical descriptions and biological interpretations of these and other concepts, the reader may refer to Barabási and Oltvai (2004) and Dong and Horvath (2007).

Random and scale-free networks are important categories of networks in the study of biological systems (Barabási, 2003). In a random network the nodes are connected to each other by chance. This also implies that the probability (or frequency) distribution of the number of edges per node follows a Poisson distribution. In a scale-free network, the probability, $P$, of the number of connections per node obeys a *power law distribution*. This distribution function can be expressed as $P(k) = k^{-\gamma}$, where $k$ represents the number of edges per node (degree) and $\gamma$ is the slope of the function [log $P(k)$]. In comparison to an exponential distribution, a power law distribution decreases more slowly. In the case of network analysis, a power law distribution describing the number of connections per node (i.e. a plot of the degree values vs. the numbers of nodes with a specific degree value) also implies that a minority of nodes in the network will be highly connected (hubs), and that the majority of the nodes will have small numbers of connections (Albert, 2005). More detailed discussions about the mathematical meaning and biological interpretations of scale-free and other types of networks can be found in (Barabási, 2003; Wagner, 2005; Csermely, 2006).

A 'network module' typically represents a group of highly connected nodes, which may be functionally similar or interrelated. A basic computational approach to detect potential modules consists of the clustering of nodes on the basis of their connectivity patterns in the network. Examples of biological-relevant network modules are protein complexes or signalling pathways. In biological systems functional modules do not act in isolation. They act in an integrated, cooperative fashion. Such an integration of modules enables a hierarchical organization of complex biological networks, that is specialized, relatively smaller modules can be combined to form larger modules with different functional roles. A comprehensive discussion about the automated detection of clusters is outside the scope of this book. Readers may refer to Rives and Galitski (2003) or Bader and Hogue (2003) for descriptions of algorithms.

## 7.3 Fundamental approaches to representing and inferring networks

Networks of gene-gene, gene-protein or protein-protein interactions can be assembled by using information (manually or automatically) extracted from the literature or by directly applying automated inference algorithms on large-scale experimental data, such as microarray data. Large sets of curated interactions can also be obtained from different public and commercial databases available on the Web (Chapter 9). Different strategies for the inference or representation of molecular interactions based on the analysis of experimental data, such as gene expression data, are introduced as follows.

Given a matrix, $\mathbf{X}$, of gene expression values, a 'gene co-expression network' can be constructed by estimating the similarity or association between pairs of genes, $\mathbf{x}_i$ and $\mathbf{x}_j$, and by graphically connecting those pairs of genes whose similarity values satisfied a pre-defined quantitative criterion. For instance, the Pearson correlation coefficient, $\mathrm{cor}(\mathbf{x}_i, \mathbf{x}_j)$, can be used to estimate similarity between the genes. And a correlation threshold value, $ct$, can be applied to establish a connection between a pair of genes in the network, that is a link is established if $\mathrm{cor}(\mathbf{x}_i, \mathbf{x}_j) > ct$. The full set of between-gene correlation values can be encoded as an adjacency or correlation matrix, of size $m \times m$, where $m$ represents the number of genes in the expression dataset.

In an 'unweighted' network the edges connecting the nodes are defined with a 'hard threshold', that is a constant threshold value, *ct*, is used for all pairs of nodes. Different methods for building weighted networks based on the application of 'soft thresholds' have also been proposed (Dong and Horvath, 2007).

Partial correlation coefficients (*parCor*) can also be computed to estimate relationships between genes and to construct co-expression networks (Schäfer and Strimmer, 2005; Keller *et al.*, 2008). In contrast to the standard Pearson's correlation, the *parCor* between two variables (e.g. genes), $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, estimates the correlation between the variables when the effects of all other variables in the dataset are considered or adjusted for. In this method, a correlation would be observed when there is an actual 'direct' relationship between the two genes, that is the correlation is not due to their possible joint dependence on a third gene. A nonzero $parCor(\boldsymbol{x}_i, \boldsymbol{x}_j)$ would mean that there is a conditional dependence between genes $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$, considering all other genes. A $parCor(\boldsymbol{x}_i, \boldsymbol{x}_j) = 0$ indicates that the two genes are conditionally independent.

Networks of interactions between transcription factors and their corresponding targets can be inferred by applying information theoretic approaches. In an information theoretic approach interactions between pairs of genes are detected by estimating their mutual information (*MI*) (Chapter 3). *MI* values above a minimum *MI* threshold would indicate an interaction between two genes. As in the case of correlation coefficient thresholds, one may define constant (hard) thresholds a priori or 'soft' thresholds using statistical analysis (Margolin *et al.*, 2006).

Different data sampling procedures, such as bootstrap (Chapter 3), can be implemented to construct and evaluate the networks inferred from data. For instance, Lim *et al.* (2009) generated 100 bootstrap datasets, which were used to infer 100 bootstrap networks. A 'consensus' network was then generated, which included the interactions most frequently detected by the different bootstrap networks.

The 'synergy' between a pair of genes with respect to a specific phenotype, for example disease or clinical outcome class, is another information theoretic measure that can be used to infer gene-gene interaction networks (Anastassiou, 2007). The synergy, $Syn(\boldsymbol{g}_i, \boldsymbol{g}_j, C)$, is defined as: $MI(\boldsymbol{g}_i, \boldsymbol{g}_j; C) - [MI(\boldsymbol{g}_i; C) + MI(\boldsymbol{g}_j; C)]$, where $\boldsymbol{g}_i$ and $\boldsymbol{g}_j$ represent the expression profiles (vectors) of the genes and $C$ represents the binary random variable encoding the presence or absence of a phenotype (Watkinson *et al.*, 2008). This measure allows the detection of potential cooperative effects or interactions of the pair of genes within a common biological process or pathway. Two genes are defined as 'synergistically' related to a phenotype if the resulting synergy is greater than zero (Watkinson *et al.*, 2008). Thus, to infer a network associated with a particular disease one can select pairs of genes whose synergy values are positive and statistically detectable. The statistical significance of synergy values can be estimated by implementing permutation tests (Chapter 2) on the gene expression data matrix.

## 7.4 Overview of key network-driven approaches to biomarker discovery

A significant diversity of approaches to discovering disease biomarkers and understanding complex relationships in the context of biological networks have been explored

(Ideker and Sharan, 2008). These approaches are based on different analysis assumptions and principles (Table 7.2).

The idea that global and local structural properties of a network can provide insights into disease-related roles of genes and proteins have been investigated using different types of networks, together with additional sources of 'omic' information. For instance, Wachi, Yoneda and Wu (2005) studied the differential expression of genes in lung cancer, and found that highly connected and central genes in a network of protein-protein interactions tend to represent genes up-regulated in cancer samples. Based on the analysis of the position of genes in networks associated with different types of cancer, Jonsson and Bates (2006) reported that genes with relevant roles in different cancers tend to show more interacting partners than non-cancer-related proteins.

Linking information from different 'omic' resources to protein networks can also provide new insights into system-based mechanisms underlying a disease or response. For instance, Goh *et al.* (2007) investigated relationships between protein interactions and gene co-expression across different diseases, which allowed them to conclude that known disease-related genes tend to encode proteins highly connected between them. Another example of network-based information integration to support the understanding of complex molecular mechanisms and responses leading to pathological conditions was proposed by Lu *et al.* (2007). They assembled a network of curated molecular interactions implicated in the allergic response in asthma. Differentially expressed genes obtained from microarray data analysis were projected onto the network to assess correlations between topology and biological functionality. Key findings of this research included the observation that highly connected network nodes tend to be less differentially expressed in comparison to nodes located in the periphery of the network (i.e. nodes of low connectivity). Moreover, they showed that potential disease-causing genes may encode neither essential nor highly connected proteins.

Camargo and Azuaje (2007) reported similar findings using gene expression data and a global protein-protein interaction network implicated in human heart failure. They found that highly connected proteins are not necessarily encoded by genes significantly differentially expressed. In addition, genes that are not significantly differentially expressed may encode a diversity of hubs and peripheral proteins. In many cases network hubs appear to be weakly correlated with interacting partners. Moreover, hubs tend to be engaged in 'higher-level' biological processes (as defined by the Gene Ontology), while peripheral nodes tend to be involved in more specific disease-related processes. These and other investigations in different biomedical domains support the notion that biomarkers with potential disease-causing roles may exhibit diverse connectivity and expression patterns. Thus, a gene does not have to be either a hub or significantly differentially expressed to represent a highly influential functional component in key processes driving health or disease.

Another important analysis principle in the analysis of disease networks is that the network neighbours of a disease-causing gene tend to have similar roles in either the same or similar disease. For example, Oti *et al.* (2006) applied this notion to the prediction of disease-associated genes in pathologies characterized by known causative genes and by additional knowledge about the potential involvement of other related genes. They showed how new disease-causing genes can be predicted by analyzing genes found in a significant genetic locus, and which encode proteins that interact with proteins encoded by genes known to be causative factors of the disease. Lage *et al.* (2007)

**Table 7.2** Representative examples of investigations of the application of network-based approaches to understanding disease and discovering novel biomarkers

| Study | Key analysis assumptions or principles | Key findings or conclusions | Reference |
|---|---|---|---|
| Differential expression of genes in lung cancer | The connectivity degree of genes in a functional network can give an indication of the disease-related roles of these genes | Highly connected and central genes in a network of protein-protein interactions tend to be genes up-regulated in cancer samples | Wachi, Yoneda and Wu (2005) |
| Analysis of the network position of genes in different cancers | Global topological properties of a network can provide insights into disease-related roles of its genes. | Genes with relevant roles in different cancers tend to show more interacting partners than non-cancer proteins | Jonsson and Bates (2006) |
| Investigation of relationships between protein interactions and gene co-expression across different diseases | Information obtained from different resources can provide new insights within a network context | Known disease-related genes tend to encode proteins that interact between them. | Goh *et al.* (2007) |
| Identification of relationships between gene expression and functional network data in human heart failure | Global topological properties of a protein-protein interaction network together with other information resources, for example gene expression and function annotation, can guide the selection of prognostic biomarkers. | Highly connected proteins are not necessarily encoded by genes significantly differentially expressed; genes that are not defined as significantly differentially expressed may encode proteins with many interacting partners; genes encoding network hubs may exhibit weak co-expression with the genes encoding their interacting protein partners. | Camargo and Azuaje (2007) |
| Prediction of disease-related genes in diseases characterized by | The network neighbours of a disease-causing gene tend to | New disease-causing genes can be predicted by looking at those | Oti *et al.* (2006) |

| | | | |
|---|---|---|---|
| known causative genes and the existence of additional incomplete knowledge about the potential involvement of other related genes | have similar roles in either the same or similar disease. | genes found in a significant genetic locus and which have protein interactions with proteins encoded by genes known to be causative of the disease. | |
| Analysis of a phenome-interactome network of protein complexes implicated in genetic disorders | Neighbouring proteins in a network display similar roles in a disease, including causative roles. | If a group of candidate proteins are members of a complex composed of proteins known to play key roles in similar diseases, then these candidate proteins will also tend be involved in a similar disease. | Lage *et al.* (2007) |
| Characterization of genes implicated in body weight based on genetic and gene co-expression network analysis | Sub-networks of interactions can represent important functional modules relevant to the emergence or progression of a disease. | Sub-networks of genes in a co-expression network can consist of many genes with strong associations with physiological traits. | Ghazalpour *et al.* (2006) |
| Analysis of a network of functional interactions relevant to breast cancer. | Sub-networks of interactions can represent important functional modules relevant to the emergence or progression of a disease. | The interconnecting neighbours of proteins known to be associated with breast cancer can represent factors relating to cancer progression. | Pujana *et al.* (2007) |
| Identification of key processes underlying cancer phenotypes using biological pathway analysis | The level of activity and relationships between genes or proteins in a sub-network can be used to distinguish between phenotypes | Gene expression data can be used to 'score' known, curated protein pathways, which can be used to predict cancer outcome. | Efroni, Schaefer and Buetow (2007) |
| Network-based classification of breast cancer metastasis | The level of activity and relationships between genes or proteins in a sub-network can be used to distinguish between phenotypes. | The expression levels of sub-networks extracted from a protein-protein interaction network can be used to predict metastasis. | Chuang *et al.* (2007) |

presented another example of the strategy of analyzing network neighbours for knowledge discovery purposes. They analyzed a 'phenome-interactome' network of protein complexes implicated in genetic disorders. Their study demonstrated that if a group of candidate proteins are members of a complex composed of proteins known to play roles in similar diseases, then these candidate proteins will also tend to be involved in a similar disease.

The idea that sub-networks of interactions can represent important functional modules relevant to the emergence or progression of a disease has also been studied. For instance, Ghazalpour *et al.* (2006) showed that sub-networks of genes in a co-expression network can include many genes displaying strong associations with specific physiological traits. This was found through the analysis of genetic and gene co-expression networks implicated in body weight. Based on the analysis of a network of functional interactions relevant to breast cancer, Pujana *et al.* (2007) found that the network neighbours of proteins known to be associated with breast cancer can also be related to cancer progression.

An extension of the previous strategy consists of the estimation of the 'level of activity' of sub-networks of genes or proteins to distinguish between phenotypes. Efroni, Schaefer and Buetow (2007) identified key processes underlying cancer phenotypes using this strategy together with extensive analysis of biological pathways. They showed how gene expression data can be used to 'score' known curated protein pathways, which in turn can be used to predict cancer outcomes. In a network-based classification of breast cancer metastasis, Chuang *et al.* (2007) demonstrated that 'expression level' measures of sub-networks extracted from a protein-protein interaction network can be used to accurately predict metastasis. The next chapter will offer a more detailed description of this and related examples of (network-driven) integrative data analysis.

## 7.5 Network-based prognostic systems: recent research highlights

The analysis of gene networks and expression profiles provided the basis for the discovery of functional mediators and pathways relevant to prostate cancer in a study reported by Ergün *et al.* (2007). This analysis was implemented with an algorithm called 'mode-of-action by network identification' (MNI), which was originally reported by di Bernardo *et al.* (2005). The MNI-based approach began by inferring a global network of gene regulatory interactions from a (training) gene expression dataset related to diverse biological processes or pathologies. This network was then used to detect genes in a test dataset, which appeared to be altered in a specific phenotype, for example disease class. This was followed by a gene ranking procedure that aimed to identify (test) genes whose expression patterns did not adequately match the inferred global network model. The rationale is that such inconsistencies may be indications of phenotype-specific perturbations on these (test) genes (Ergün *et al.*, 2007).

In Ergün *et al.* (2007) the training dataset used by the MNI algorithm comprised 1144 microarray expression profiles originating from 13 research projects in different cancer areas, such as breast and prostate cancer. The test dataset consisted of expression data from prostate cancer samples only: non-recurrent primary vs. metastatic samples. The samples in this test dataset were then queried against the network inferred by the MNI algorithm on the training dataset. The idea was to find genes whose expression patterns did not fit the global model of expression regulation extracted from multiple cancer types. This allowed the researchers to focus on genes that can be defined as prostate cancer-specific candidates. Thus, those genes representing 'inconsistent' profiles in relation to the global network were retrieved and ranked. This (test) gene expression-network matching and ranking process was carried out within each prognostic category, that is potential biomarkers were retrieved from the two patient groups independently. This approach enabled the authors to propose the androgen receptor gene (AR) and the AR pathway as potential key mediators of metastasis in prostate cancer.

Watkinson *et al.* (2008) applied the concept of synergy of gene pairs to estimate their involvement in a specific disease (Section 7.3). In this approach a pair of genes displaying a high synergy (with respect to a specific phenotype) indicates that these genes participate in a common biological pathway associated with the phenotype. Thus, synergy can be used as an estimator of the functional relation of two genes, or of their 'cooperative' effects, in the context of a specific phenotype, such as the presence of disease.

Lim, Lyashenko and Califano (2009) discovered potential upstream 'master regulators' of breast cancer, which were shown to provide better predictive power and robustness than other expression signatures when tested on both their original and independent datasets. This is especially significant given the difficulties of previous studies, including works published in prestigious journals, to meet the challenges of independent (follow-up) validation studies. Moreover, this also addresses the problem of biomarker set instability, that is the little agreement between sets of biomarkers found in different studies relating to the same clinical problem. The latter can be observed in two studies that identified gene expression signatures in breast cancer patients to predict progression to metastasis (van de Vijver *et al.*, 2002; Wang *et al.*, 2005), but which shared only one gene in common.

The ideas explored by Lim, Lyashenko and Califano (2009) and others in relation to the discovery of 'driver' biomarkers can be explored by applying algorithms that infer, for example, transcriptional factors from gene expression data. The ARACNe algorithm is one example of such algorithms (Basso *et al.*, 2005), which have been experimentally validated using different types of cells. In Lim, Lyashenko and Califano's study (2009), ARACNe detected transcription factors involved in induction or suppression of genes associated with differential prognosis in breast cancer, based on the analysis of two published datasets (van de Vijver *et al.*, 2002; Wang *et al.*, 2005). The network obtained from one of these datasets (van de Vijver *et al.*, 2002) was interrogated by a Master Regulator Analysis (MRA) procedure. Given two phenotypes, the MRA identifies transcription factors that may be used to explain the differences observed in the two phenotypes. The main outcome of the MRA is a set of phenotype-specific 'master regulators' ($MR_{\mathrm{i}}$) and corresponding regulons or transcriptional targets ($R_{\mathrm{i}}$). The MRA tests whether the regulons activated by the transcription factors are significantly represented in the genes over-expressed in one of the phenotypes (e.g. the samples in the case group). Similarly, it tests whether the genes inhibited by the transcription factors are statistically overrepresented in the genes that are down-regulated in the case group. In Lim, Lyashenko and Califano's study (2009), the set of $MR_{\mathrm{i}}$ genes were subsequently evaluated as inputs to support vector machine classifiers using the two datasets independently. It should be noted that the sets of regulons found with this analysis overlapped with the set of biomarkers derived from the original study by van de Vijver *et al.* (2002).

Based on the analysis of 'pathway activity levels' extracted from each patient, Lee *et al.* (2008) proposed a classification method based on the mapping of gene expression data onto different biological pathways, which was shown to outperform traditional methods in different disease classification applications. They defined the 'activity level' as a measure that summarizes the levels of gene expression of 'condition responsive genes' (CORGs) found in a biological pathway. In a given pathway, the CORGs are the sub-set of genes that allow an optimal discrimination of the phenotypes studied based on their combined expression values. Thus, the resulting biomarkers represent sub-sets of functionally-related, phenotype-specific genes. This method is described as follows.

The investigated pathways were obtained from the MsigDB resource (Subramanian *et al.*, 2005): 472 metabolic and signalling pathways derived from 8 manually annotated databases and 50 clusters of co-expressed genes originating from different investigations. For a given gene set, **G**, within a particular pathway, the discriminative score $\boldsymbol{S}(\mathbf{G})$ is equal to the t-test statistic obtained on **G** for the phenotype classes considered. $\boldsymbol{S}(\mathbf{G})$ is actually based on the t-statistic obtained on the average expression values of the genes in **G**(Lee *et al.*, 2008). Other authors have proposed alternative methods to estimate the activity of a pathway based on principal component analysis (Bild *et al.*, 2006), means and medians (Guo *et al.*, 2005), and gene-class correlation values (Tian *et al.*, 2005).

For each pathway, Lee *et al.* (2008) mapped the expression values of each gene onto its corresponding gene (protein) in the pathway, and searched for a sub-set of genes that could be used to differentiate between the samples of the phenotypes investigated. The output of this search is a set of genes with a maximal $\boldsymbol{S}(\mathbf{G})$ in the pathway. These genes in turn represented the CORGs. These CORGs were identified by ranking the t-test statistic values of the different candidate sub-sets. The initial member of a CORG set, **G**, is the gene with the largest t-test score in the pathway. In subsequent iterations, the gene with the next largest t-test score is added to **G**. The search is ended when, for a new gene addition, $\boldsymbol{S}(\mathbf{G})$ cannot be increased.

Another approach to network-based classification of phenotypes or clinical outcomes consists of looking at structural changes in interaction networks that can be linked to functional properties. This can include the detection of inter- or intra-module hub proteins that are co-expressed with interacting partners in specific cells, tissue types or phenotypes. In this scenario, for instance, the modification of network modularity can be associated with specific clinical outcomes for patient classification. Taylor *et al.* (2009) applied this strategy by first identifying hubs in networks of annotated protein-protein interactions extracted from the literature and different experimental resources. Using large-scale gene expression data, they estimated the level of co-expression of these hubs and their interacting partners. Such a co-expression analysis helped them to define inter- and intra-modular hubs. Inter-modular hubs tend to display low co-expression with their interacting partners. Intra-modular hubs exhibit stronger patterns of correlation with their interacting partners.

With the analysis of these network and expression data, Taylor *et al.* (2009) concluded that intra-modular hubs are more functionally similar with their interacting partners in comparison to inter-modular hubs. Also their results indicated that inter-modular hubs tend to reflect tissue-specific molecular mechanisms. Thus, their research showed that these types of hubs exhibit structural features that are linked to their functional roles in organizing and modulating protein networks. Moreover, they found that mutations of inter-modular hubs are associated with cancer phenotypes, and that such associations are more frequent in comparison to those found in intra-modular hubs (Taylor *et al.*, 2009).

Taylor *et al.* (2009) also examined the statistical differences in the average (Pearson) correlation coefficients displayed by hub proteins and their interacting partners in two groups of cancer patients: good and poor clinical outcome. This reported 256 hubs with statistically detectable changes in correlation values across good and poor outcome. A closer look at these results suggested, for instance, that the loss of coordinated co-expression of groups of genes associated with the BRCA1-associated genome surveillance complex (BASC) can be a driving mechanism leading to poor clinical outcome (Taylor *et al.*, 2009). Automated classification of patients was then implemented as follows. First, they estimated the co-expression of hubs with their interacting partners. Second, they identified hubs with expression correlations that significantly differed between the prognostic classes (survival vs. death groups). Using this information they implemented classification models based on an algorithm known as 'affinity propagation clustering' (Frey and Dueck, 2007). The goal was to predict 10-year survival of patients. Through a fivefold cross-validation procedure and ROC curves (Chapters 2 and 3), they demonstrated that their network-based classification strategy can outperform commercial prognostic systems based on gene expression data only.

## 7.6 Final remarks: opportunities and obstacles in network-based biomarker research

The knowledge extracted from different types of molecular networks can provide the basis for the discovery of novel biomarkers, functional pathways and processes, which can guide the implementation of more meaningful diagnostic, prognostic and treatment response prediction systems. In different health and pathological conditions it has been shown, for instance, that proteins encoded by genes with disease-related genomic mutations can be grouped into common pathways, complexes or processes. This has been demonstrated through the analysis of experimentally- or computationally-inferred networks. Also it is known that gene products functionally associated with a specific phenotype may be found as interacting partners in clusters of network nodes, which can represent protein-protein interactions or co-expression relationships.

The comparative analysis of protein networks derived from humans and pathogens is a promising approach to understanding infection and defence mechanisms (Ideker and Sharan, 2008). The identification of interactions and pathways distinguishing pathogens (i.e. viruses, bacteria or parasites) and hosts can facilitate protein target identification and drug development. Other important pharmacological applications in other biomedical areas involve the analysis of networks describing chemical-genetic interactions, and protein-protein interaction networks integrated with drug-drug interaction networks (Ideker and Sharan, 2008).

The difficulties in identifying robust gene expression biomarkers may be explained by differences and quality issues regarding sample extraction and processing, microarray technologies and data analysis. Nevertheless, these factors do not fully explain this lack of stability and reproducibility. Lim, Lyashenko and Califano (2009) argued that the main reason is that most prognostic and diagnostic signatures discovered to date represent 'passengers', rather than 'drivers' of disease.

This explanation is supported by the notion that most of the biomarkers reported in the literature tend to represent highly-differentially expressed genes, which in turn tend to represent genes located 'downstream' from the primary (somatic or inherited) factors driving the prognostic outcomes. Furthermore, because of the large and complex interrelationships defined by regulatory networks, such 'downstream' genes tend to be unstable or highly variable in terms of expression patterns. The higher the complexity of this regulatory interplay, the larger the number of co-regulators and noise influencing the processes causing the differential expression observed in the 'passenger' genes. Lim, Lyashenko and Califano (2009) and others have also pointed out that this can be seen, for example, in the case of oncogenes and tumour suppressors, which do not tend to be the most differentially expressed genes.

This makes network-based approaches a promising solution to support the discovery of 'driver' agents of disease. This may comprise, for instance, genes and products responsible for initiating the cascade of transcriptional responses leading to differential expression patterns across phenotypes or patient groups.

The sparseness and incompleteness of available human network information will continue to be one of the major challenges for the advancement of network-based approaches to biomarker discovery and beyond. There is also a need to accumulate more, better quality and less biased datasets describing different types of interactions in different clinical conditions, tissues or cell types. Moreover, there is still a need for alternative computational methodologies and tools to collect, assemble, visualize and analyze these networks in the context of translational biomedical research.

Many of the network-based approaches published to date are based on the inference of structural properties and relationships. It has been shown that the knowledge of the interactome of humans and other organisms not only remains incomplete, but also biased (Hakes *et al.*, 2008). This represents a crucial problem, despite advances in experimental approaches and the increasing amounts of data deposited in the literature and curated databases. Biases may be caused by differences in the data acquisition and processing procedures of these datasets. These biases can distort the structure of a network. Therefore, it is necessary to consider both the incompleteness of knowledge and potential sources of bias when making interpretations based on the topology of molecular networks. Hakes *et al.* (2008) argued that researchers generally accept that molecular networks can miss large numbers of interactions, and that many of the reported interactions may represent false positives. However, problems related to data sampling bias have received relatively less recognition. Interactome networks can be biased toward proteins from similar cellular environments and toward highly expressed proteins. Moreover, these networks may be biased toward the study of proteins that are more ancient or more evolutionarily conserved. In the case of high-quality collections of molecular interactions, another potential source of bias is that proteins included in curated interactions tend to be the ones that have been more studied as potential disease-related or essential proteins.

Hakes *et al.* (2008) explained that current interaction networks are rough approximations or samples of the 'complete' networks investigated. They also asserted that even if this sampling was truly random, the incompleteness of the resulting networks would be significant. Additionally, sampling biases would generate larger differences between the "complete network" and the sub-sample network actually analyzed. Using different published interactome datasets, Hakes *et al.* (2008) demonstrated that global statistical descriptors and topological properties can be significantly affected by data handling and selection. The authors recommend caution when making conclusions about the structure of molecular networks and that more attention should be given to the identification of potential sources of biases.

Another area that deserves greater attention is the study of the tissue specificity of protein interaction networks. Most of the human interaction networks reported to date represent approximations of 'global' interactome networks. Such networks include very limited information about where and when the interactions occur. The analysis of dynamic interaction networks in unicellular organisms has been studied through the integration of protein-protein interaction and gene expression data in the context of different cellular states or conditions. Examples of this research in yeast are the identification of co-regulated interaction modules (Ihmels *et al.*, 2002) and the investigation of cellular conditions under which interactions occur (Luscombe *et al.*, 2004). More recently, Bossi and Lehner (2009) identified human tissues in which different protein interactions can occur based on the analysis of gene expression data. One of the key conclusions of their study was that there are abundant interactions between proteins that are globally expressed and those proteins that appear to be expressed in specific tissues only. For instance, they showed that most tissue-specific proteins tend to interact directly with proteins implicated in fundamental cellular processes. Moreover, 'housekeeping' proteins also tend to have many tissue-specific interactions (Bossi and Lehner, 2009). This and future research about the dynamic behaviour of protein interaction networks in the context of different tissue types will be required to have a deeper understanding of the mechanisms underlying health and disease, and to guide the search for potential biomarkers.

# Guest commentary on chapter 7: Commentary on 'disease biomarkers and biological interaction networks'

**Zhongming Zhao**
*Departments of Biomedical Informatics, Psychiatry and Cancer Biology Vanderbilt University School of Medicine, Nashville, TN 37203, USA*

Discovery of novel molecular biomarkers of disease has recently become a major research topic in translational research thanks to the coming of the genomic era. Because of its biological and economical importance, biomarker discovery has been strongly supported by US government agencies as well as funding agencies in other countries. In the blueprint for the genomic era proposed by Francis Collins and colleagues in the US National Human Genome Research Institute (NHGRI) in 2003, most of the grand challenges in the applications of genomics to health are much related to biomarker discovery, including identification of disease causal markers at the gene, locus, network and pathway levels, prediction of disease susceptibility and drug response, early

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

detection of illness, and development of powerful new therapeutic approaches to disease (Collins *et al.*, 2003). As described in Chapter 7, network-based approaches using different types of 'omic' data or their combination have been extensively applied to the biomarker discovery in many diseases such as cancers, cardiovascular diseases, and diabetes. The 'omic' data that have been recruited in biomarker discovery includes protein-protein interactions, protein-gene interactions, gene expression and co-expression, disease gene and phenotype relationships, and gene-environment interactions. These biological datasets have been generated during the past two decades by taking advantage of the revolutionary high-throughput technologies of genomics, epigenomics, transcriptomics, and proteomics, and the wide-spread proliferation of biology-oriented databases and computational mining tools. Despite the well known incompleteness and high false positive rate in the 'omic' datasets, analyses using these datasets by systems biology approaches have provided many important insights into the molecular mechanisms of diseases.

Here I discuss three current trends in biomarker discovery in complex disease. First, complex diseases such as cancers and psychiatric disorders are likely caused by many genes, each of which may contribute a small risk but likely interact with other genes and/or interact with environmental factors. It is now clear that disease prediction by using individual biomarkers is limited in many cases (Rifai, Gillette and Carr, 2006). Alternatively, a panel of genes or proteins is needed to accurately detect the perturbation of the biological systems that cause the disease. At present, there is a strong trend towards integrating the data from various genetic studies and their related biological information in the cellular systems so that promising candidate biomarkers can be screened by enriched evidence and at the systems biology level. Second, investigators are applying pathway-based analysis for the enriched disease-causal information from independent large-scale or genome-wide genetic data, such as emerging genome-wide association studies (GWAS). The third trend is to integrate the disease-related pathways and networks (i.e. pathway crosstalk) for complex disease studies. More details are discussed below.

## Integrative approaches to biomarker discovery

A tremendous amount of effort has been expended in the past two decades to identify genes influencing susceptibility in complex diseases, such as schizophrenia and Alzheimer's disease. The identification of potential complex disease susceptibility genes is expected to accelerate because of many GWA studies using large datasets currently complete or in progress. Concurrently, across a variety of complex disorders and traits, there is a strong trend towards the integration of data from multiple sources and the use of these integrated data to generate lists of prioritized candidate biomarkers at the systems level. This strategy has recently been effectively applied in the discovery of candidate biomarkers in many diseases such as cancers and major psychiatric disorders (Lin *et al.*, 2007; Le-Niculescu *et al.*, 2009). For example, the Niculescu group in Indiana University developed a convergent functional genomics (CFG) approach, which integrates functional (e.g. gene expression), genetic (e.g. linkage and association studies), and tissue and fluids (e.g. blood, postmortem brain) data, and then applied a Bayesian strategy for cross-validation and prioritization of biomarker genes. They applied this strategy in almost all major psychiatric disorders such as schizophrenia, bipolar disorder, alcoholism, and mood disorder. In their most recent study, they identified blood biomarkers for mood disorders. These biomarker genes were found to be involved in myelination and growth factor signalling pathways (Le-Niculescu *et al.*, 2009). Similarly, Sun *et al.* (2009) developed a multi-dimensional evidence-based gene prioritization approach for complex diseases. To demonstrate this approach, Sun *et al.* integrated evidence-based genetic data from thousands of association studies, more than 25 genome-wide linkage scans, meta-analysis of gene expression and high-throughput literature search. The prioritized candidate genes were then used to construct gene networks. Subsequent analysis identified a few small schizophrenia-specific sub-networks that are enriched in genetic signals from independent GWA studies. The follow up experiments verified that some of the genes in the sub-networks are significantly associated with schizophrenia. These sub-networks provide candidate network-based biomarkers for schizophrenia.

One example of the application of integrative approaches to cancers was described in Lin *et al.* (2007). In that study, multi-dimensional integrative analysis based on sequence similarity, functional annotations, protein-protein interactions, and molecular pathways was performed to identify functional groups and pathways that are enriched for mutations relevant to both breast and colorectal cancers. The framework provides an efficient approach for biomarker discovery in cancers.

With many genes having been found associated with different complex diseases and the recent availability of the human interactome and whole molecular networks (e.g. constructed by the Ingenuity Pathway System), investigators have started to construct disease-specific networks. For example, Jin *et al.* (2008a) constructed a prostate cancer-related network (PCRN) by searching for prostate cancer genes in the Ingenuity Pathway Systems and protein-protein interactions in the Human Protein Reference Database (HPRD). We recently constructed a schizophrenia-related molecular network by combining the sub-networks and pathways in which schizophrenia candidate genes are known to be involved. This is the first molecular network model of psychiatric genetics. In cardiovascular diseases, Jin *et al.* (2008b) built a cardiovascular-related network using an integrated knowledge, network-based biomarker discovery scheme. In their scheme, they integrated data from mass spectrometry (MS) and network and pathway (Uniprot, KEGG, and HPRD) data. They demonstrated that candidate network-based biomarkers can be more accurate in classifying different groups of patients than single biomarkers.

## Pathway-based analysis of GWA data

Genome-wide association studies test many thousands to more than one million markers at a time. They are unbiased, hypothesis-free, and aimed at the discovery of novel disease-casual variants. Under the hypothesis that many genes contribute a small risk to complex disease, the detection of single biomarkers at the genome-wide significance level is often challenging, especially for neuropsychiatric disorders. In schizophrenia, the only published GWA study failed to identify genes to be significantly associated with schizophrenia at the genome level. However, pathway-based analysis of GWA data is emerging as a useful tool. It assumes that markers underlying a disease or phenotype are enriched in genes belonging to the same pathway. One popular pathway analysis method for GWA data is based on the gene-set enrichment analysis (GSEA) algorithm originally developed for microarray data analysis (Wang, Li and Bucan, 2007). It defines sets of genes based upon common biological attributes (e.g. Gene Ontology terms or biological pathways) and measures the degree of overrepresentation or ‘enrichment’ of each gene set amongst nominally disease-associated markers. So far, pathway-based analyses have been quickly applied to many diseases such as Parkinson disease, Axon guidance, multiple sclerosis, bipolar disorder, and Crohn disease. All these studies could identify biologically important pathways related to the corresponding disease.

Pathway-based analysis can be applied to other genome-wide data too. For example, a genome scan meta-analysis (GSMA) of the 32 genome-wide linkage studies for schizophrenia has been just completed. A follow-up pathway analysis of the resulting candidate genes identified myelin-related pathways implicated in schizophrenia (Rietkerk *et al.*, 2009).

## Integrative analysis of networks and pathways

Given the complex nature of biological systems, more than one pathway may be involved in any given complex disease. Two or several pathways or networks may interact with each other to cause the disease. This is very likely because functional important proteins (e.g. TP53) may be involved in multiple pathways. Therefore, besides the identification of specific pathways/networks, investigators may take a further step by exploring the interaction and crosstalk between pathways that are related to a disease. This integrative or pathway crosstalk analysis has been applied in cancer genes (Lin *et al.*, 2007; Li, Agarwal and Rajagopalan, 2008). We recently applied this approach to schizophrenia. We first identified 24 schizophrenia-related pathways. Then, we developed a statistical method to evaluate whether the schizophrenia-related proteins (nodes) and their interactions (links) were significantly shared between any pair of schizophrenia-related pathways. Based on the pathways that are shared, we constructed a network of crosstalking pathway. We found that neurotransmitter-related pathways are strongly in crosstalk, which suggest the neurotransmitter hypothesis of schizophrenia.

## References

Collins, F.S., Green, E.D., Guttmacher, A.E. and Guyer, M.S. (2003) A vision for the future of genomics research. *Nature*, **422**, 835–847.

Jin, G., Zhou, X., Cui, K. and Wong, S.T.C. (2008a) The network biomarker discovery in prostate cancer from both genomics and proteomics levels. OSB’08, pp. 144–151. ORSC & APORC, Lijiang, China.

Jin, G., Zhou, X., Wang, H. *et al.* (2008b) The knowledge-integrated network biomarkers discovery for major adverse cardiac events. *Journal of Proteome Research*, **7**, 4013–4021.

Le-Niculescu, H., Kurian, S.M., Yehyawi, N. *et al.* (2009) Identifying blood biomarkers for mood disorders using convergent functional genomics. *Molecular Psychiatry*, **14**, 156–174.

Li, Y., Agarwal, P. and Rajagopalan, D. (2008) A global pathway crosstalk network. *Bioinformatics*, **24**, 1442–1447.

Lin, J., Gan, C.M., Zhang, X. *et al.* (2007) A multidimensional analysis of genes mutated in breast and colorectal cancers. *Genome Research*, **17**, 1304–1318.

Rietkerk, T., Boks, M.P., Sommer, I.E. *et al.* (2009) Network analysis of positional candidate genes of schizophrenia highlights myelin-related pathways. *Molecular Psychiatry*, **14**, 353–355.

Rifai, N., Gillette, M.A. and Carr, S.A. (2006) Protein biomarker discovery and validation: the long and uncertain path to clinical utility. *Nature Biotechnology*, **24**, 971–983.

Sun, J., Jia, P., Fanous, A.H. *et al.* (2009) A multi-dimensional evidence-based candidate gene prioritization approach for complex diseases – schizophrenia as a case. *Bioinformatics*, **25** (19): 2595–2602.

Wang, K., Li, M. and Bucan, M. (2007) Pathway-based approaches for analysis of genomewide association studies. *American Journal of Human Genetics*, **81**: 1278–1283.

# 8 Integrative data analysis for biomarker discovery

This chapter focuses on the combination of different types of information and prediction models for biomarker discovery. The importance and application of integrating different types of data and computational approaches will be discussed. Methodologies and tools for integrating and analyzing different data sources will be introduced. Examples of approaches to supporting the identification of biomarkers for disease classification and the prediction of clinical outcomes will be provided.

## 8.1 Introduction

The combination of multiple biomarkers derived from different clinical and molecular data sources have been proposed to improve diagnostic and prognostic performance in different research areas, especially in cancer and cardiovascular research. For example, traditional risk factors (such as age, gender and blood glucose concentration) have been combined with protein expression biomarkers (such as different molecules implicated in inflammation processes) to improve the prediction of recurrent cardiovascular events in comparison to traditional risk factors (Blankenberg *et al.*, 2006).

In a traditional integrative approach to disease prediction model design, the different biomarkers are independently discovered prior to the integrative modelling task. Typically they are selected as inputs to these analyses based on prior knowledge, that is new models are investigated based on the combination of known traditional risk factors or novel molecular biomarkers, whose computational predictive quality and potential

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

clinical relevance have already been reported elsewhere. This is the case of the example pointed out above. An alternative integrative approach in the 'omic' era comprises the combination of different data and information sources or computational techniques to discover and prioritize new biomarkers, as well as to implement novel prediction methodologies. Moreover, under this framework the prediction model discovery process does not have to be a priori biased by strong user- or domain-driven assumptions about between-biomarker or biomarker-outcome relationships. This chapter puts emphasis in this latter strategy based on different types of 'omic' datasets and information, computational learning and statistical techniques, and knowledge bases.

A key rationale for model integration is that different prediction models, such as classifiers, built on different data sources or using different techniques can potentially provide complementary information about the data to be classified. This capacity to offer complementary or alternative vistas of the same problem also means that the different subsets of misclassified or incorrectly predicted samples by these models will overlap partially. The main idea behind the integration of data and models for biomarker discovery and disease classification is to achieve consensus or better predictions based on the combination of several 'opinions' provided by different 'experts'. These experts can be seen as relative weak predictors or descriptors of the problem under consideration. Their individual assessments are also based on incomplete, noisy and often ambiguous information.

The areas of computational intelligence and data mining have contributed different integration schemes, algorithms and applications that have been shown to outperform models based on single data sources in a wide variety of problem domains. An important condition highlighted in these investigations is that the data sources and resulting models to be integrated should be as diverse as possible (Kittler *et al.*, 1998). In practice, this can be achieved by building models using different types of clinical and molecular data representing different levels of 'omic' complexity or organization, by using different feature sets or by sampling different training datasets. Another strategy consists of extracting predictions from specific models according to their application context or prediction capability in relation to local constraints, class-specific conditions or data subsets (Kittler, 2000; Hastie, Tibshirani and Friedman, 2001).

A variety of integration strategies based on different types of data and machine learning techniques have also been studied for the implementation of 'ensembles' of classifiers (Kittler *et al.*, 1998). For instance, majority or weighted voting can be implemented to generate integrated predictions in situations when only class labels are available. When the predictions are defined by continuous values, for example probabilities or risk scores, linear combinations and mean values can be calculated to produce the integrated predictions. Model integration applications may comprise different single models built on a single dataset or different datasets encoded by a common input representation or feature set format, such as different gene expression datasets from different populations but measured on the same genes. In this case predictive diversity may be accomplished by generating models based on different algorithms (e.g. support vector machines, neural networks and instance-based learners) or based on the same technique with different designs (e.g. neural networks trained with different learning parameters, topologies, architectures or evaluation requirements). Another typical scenario in the implementation of integrated prediction models is to build diverse models trained on different datasets assumed to be independent or weakly correlated

(Kittler, 2000), which may also be derived from different types of data. An example is the integration of gene expression and genomic variation datasets originating from the same group of patients.

In general, one may define five major types of approaches to data and prediction model integration for biomarker discovery research on the basis of the integration strategy implemented. In the first category models are constructed based on the aggregation, such as union or intersection, of different features at the input level. In this case the integration is done before any classifier or prediction model is trained. In a second family of approaches different models are generated from a single dataset or homogeneous data sources, such as a single gene expression dataset or multiple expression datasets measured on the same genes, followed by the integration of the resulting models to obtain fused or global prediction outcomes. A third category involves the integration of different or heterogeneous data sets during the construction of the prediction models. This can be done, for example, as part of data pre-processing or feature encoding tasks. Another family of approaches consists of combing different heterogeneous datasets, information sources and the resulting prediction models in a parallel fashion. Examples of approaches assigned to this category also include the combination of heterogeneous 'omic' data types or datasets annotated to different, but inter-related, phenotypes. The fifth major methodological category includes the multi-stage, serial integration of multiples datasets or prediction models. These major categories are graphically illustrated in Figure 8.1.

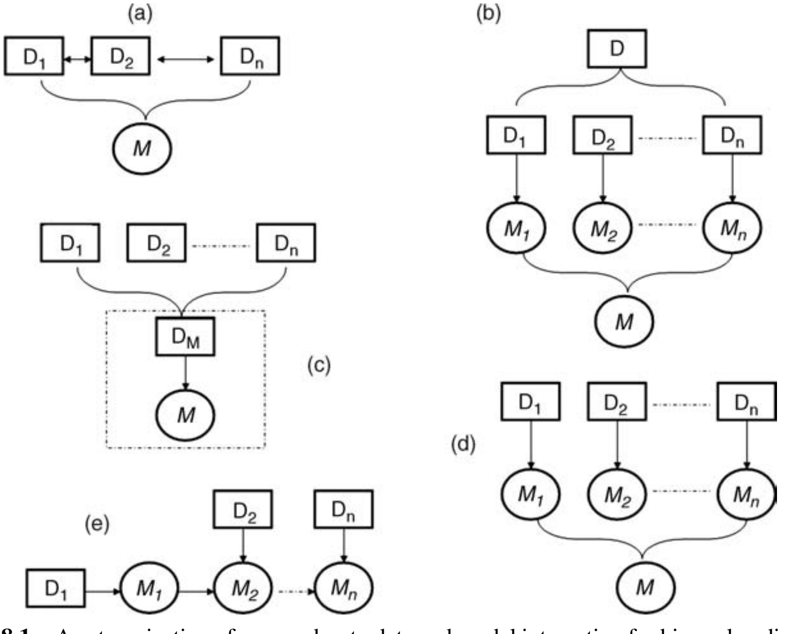


**Figure 8.1** A categorization of approaches to data and model integration for biomarker discovery and prediction on the basis of the integration strategy implemented. (a) Model based on data source aggregation at the input level; (b) integration of models based on homogeneous data sources; (c) data integration at the model level; (d) multiple heterogeneous data and model integration; (e) serial integration of sources and models. *D*: data or information source. *M*: Prediction model. Arrows represent transfer of information

A categorization of approaches to data and prediction model integration in disease biomarker discovery research can also be defined on the basis of the outcome representation or visualization methodology applied. In general, one can identify two major types of strategies: Component- and network-centric strategies. A component may refer to any single biological or clinical concept, such as a protein or gene biomarker. In the network-centric approach the resulting outcomes, models or predictions are visualized as networks of components. A network can be associated with different levels of complexity, such as a pathway or other functional interaction network (Chapter 7). Nodes may represent different types of biological system components. Edges within such networks can represent different types of functional relations, such as regulatory or protein-protein interactions. The component- and network-centric strategies are graphically depicted in Figure 8.2.

The next sections will discuss these strategies in more detail with relevant examples and in the context of biomarker discovery, prioritization and biomedical knowledge discovery. Although these categories can partially overlap or it may be possible to assign some of these examples to different approach categories, the main objective is to highlight the most representative design and application attributes that link these examples to a particular category.

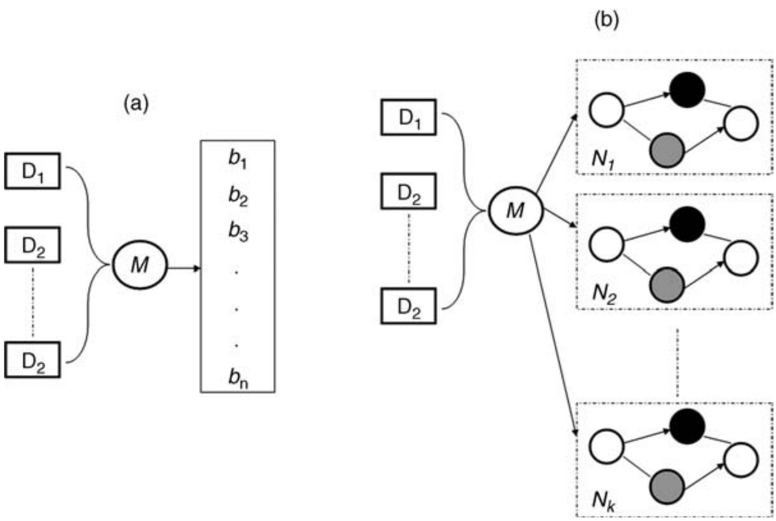


**Figure 8.2** A categorization of approaches to data and model integration for biomarker discovery and prediction on the basis of the prediction encoding and visualization implemented. (a) Component-centric approach; (b) Network-centric approach. Component refers to any single biological or clinical concept, such as a protein or gene biomarker. *D*: data or information source. *M*: Prediction model. *N*: Network-based representation of resulting outcomes, models or predictions. A network can be associated with different levels of complexity, such as a pathway or different functional interaction networks. Nodes may represent different types of biological system components. Edges within such representations can represent different types of functional relations, such as regulatory or protein-protein interactions

## 8.2 Data aggregation at the model input level

This category includes prediction models that process inputs representing diverse features or prediction factors. The features can represent different measurements, biomarkers and types of molecular and clinical data. This can also mean that a patient is represented by a set of predictive features measured with different experimental protocols or instruments, but which are aggregated into a single input, for example a vector of values, to make predictions. Within this category, we can also include classification models based on biomarkers previously discovered in different studies independently. This in turn means that new measurements of such biomarkers are made on a specific patient cohort to construct the new 'integrated' prediction models.

Blankenberg *et al.* (2006) illustrate a typical example in which known biomarkers of cardiovascular disease: nine biomarkers of inflammation, microalbuminuria and N-terminal pro-brain natriuretic peptide (NT-proBNP), were measured in more than 3000 individuals. The main goal was to detect significant associations between sub-sets of these genes and different cardiovascular outcomes, such as myocardial infarction and death, over 4.5 years of follow-up. The research concluded that the combination of NT-proBNP and traditional risk factors (e.g. age, cholesterol and glucose levels) could improve the clinical outcome prediction capacity in comparison to models based on traditional biomarkers only.

The automated selection and prioritization of known diagnostic and prognostic biomarkers has also been investigated by other authors, such as Mamtani *et al.* (2006). In the context of binary diagnostic classification and using different types of 'omic' data, they aimed to demonstrate how predictive performance could be improved based on the combination of the most powerful biomarkers discovered in previous investigations. The proposed methodology can be summarized by the following sequence of analysis phases. In a first phase, the AUC value (Chapter 2) from each biomarker was estimated independently. This information represented the basis for the calculation of a 'performance index', which was used to compare the individual predictive quality of the different biomarkers. In the second phase, and based on the set of top biomarkers, an optimized sub-set of biomarkers was selected by stepwise multiple linear regression (Chapters 2 and 3). This sub-set of biomarkers represented the inputs to the third phase: the implementation of a classification model based on a linear discrimination function, and its subsequent validation.

The automated prioritization of biomarkers has also been researched by Aerts *et al.* (2006) and others, but in the context of multiple, heterogeneous information sources. Moreover, unlike Mamtani et al.'s (2006) approach, Aerts *et al.* (2006) developed a Web-based system that can be applied to prioritize (potential novel) candidate biomarkers. Section 8.5 offers a more detailed description of this example.

## 8.3 Model integration based on a single-source or homogeneous data sources

Prediction models trained on extended datasets assembled through the combination of samples derived from independent studies, including those publicly available on

Web-based resources (Chapters 9 and 10), can potentially improve diagnostic and prognostic applications in comparison with models based on single datasets generated at a single laboratory. Also, it is known that different models built on independent datasets (e.g. several gene expression datasets only) can incorporate different sets of biomarkers with little overlap between them. Moreover, models based on different biomarker sets can show similar classification performance when tested on the same (independent) dataset (Ein-Dor *et al.*, 2005; Fan *et al.*, 2006; Zhang *et al.*, 2007a).

Different (known and unknown) confounding factors in each dataset also represent a major problem for the predictive integration of these datasets. Examples of typical confounding factors are lymph node status in cancer patients, diabetes or metabolic complications in control or case groups in cardiovascular research, and differences between datasets in terms of ethnic or bio-geographical origins. This means that the biomarker information obtained from a dataset may not be directly applicable to or harmonized with other datasets that include patients with diverse demographical and clinical backgrounds. Thus, an important task prior to the integration of samples originating from independent studies is to detect and control for potential confounding factors. Standard strategies to deal with this challenge include the design of prediction models specific to sub-populations and the inclusion of large, diverse datasets (Zhang *et al.*, 2007a). These problems may equally apply to some applications based on other integrative approaches, such as the integration of multiple heterogeneous datasets.

A study by Zhang *et al.* (2007a) addressed the problem of integrating independent microarray datasets for building prognostic models. The classification problem was the prediction of outcomes in breast cancer patients (disease recurrence vs. good prognosis) using published microarray data. A lack of clinical information or metadata about the microarray datasets, as well as the lack of evidence about potential sources of bias in each dataset, did not facilitate the detection of confounding factors prior to the integration of the samples into a single gene expression dataset. Zhang *et al.* (2007a) created training datasets using the samples obtained from two independent microarray datasets, but which represented the same classification problem. The integration was performed after implementing standard expression normalization within each dataset independently: the median and standard deviation of the expression values for each gene were made equal to zero and one respectively. The resulting prediction models were subsequently tested on a third independent dataset. This relatively simple integration approach produced prognostic models of higher overall classification accuracy in comparison to models built on single datasets.

This category also includes data mining approaches to identify potentially relevant relationships or data patterns for supporting disease biomarker discovery. Such relationships can be used, for instance, as inputs to subsequent prediction model design tasks or to reduce the cost and complexity of future data acquisition and analysis. An example of this type of approach was reported by Alterovitz *et al.* (2008), who integrated protein expression data obtained from different types of bio-fluids and tissue samples. This investigation allowed them to detect direct associations between tissue proteins and unknown counterparts measured in peripheral fluids, that is it determined which fluid biomarkers might be used as proxies to tissue biomarkers. This knowledge is particularly important because peripheral or circulating biomarkers, such as those measured in blood or urine, are easier and cheaper to obtain than those extracted from solid tissue or organ biopsies. Associations between solid tissues and fluids were estimated by calculating their ‘relative entropy’ (Hastie, Tibshirani and Friedman, 2001) in the context of phenotypes, functions and drugs. Relative entropy allows one to estimate the similarity between two distributions or datasets, in this case: tissues vs. fluids. Lower entropy values indicate higher similarity between the two sources. In the case of Alterovitz *et al.*’s study (2008), a set of fluid biomarkers was considered informative or representative of a set of solid tissue biomarkers when their relative entropy scores were significantly greater than the score obtained from randomly-selected biomarker sets. The latter including the same number of biomarkers as found in the original set of fluid-derived biomarkers. Relative entropy values were estimated using the frequency of functional annotation terms found in the biomarker sets under comparison. Thus, this type of analysis enables the identification of inter-source commonalities beyond the mere identification of ‘overlapping’ biomarker sets. Alterovitz *et al.* (2008) carried out this integrative data analysis on 16 solid tissue and 10 fluid (proteome) datasets. Functional information of the proteins was extracted from the GO, and the Online Mendelian Inheritance in Man database (OMIM, 2009) was used to map proteins to diseases. Protein-drug associations were derived from an ontology based on the Pharmacogenetics and Pharmacogenomics Knowledge Base (PharmGKB) (Hewett *et al.*, 2002).

The integration of multiple microarray datasets for the discovery of potential novel biomarkers has also been investigated in the context of protein-protein interaction networks, extensive literature data mining and cardiovascular research. Camargo and Azuaje (2008) demonstrated the predictive capability of classification models inferred from such an integrative analysis process, in comparison to single-source prognostic systems based on biomarkers known to be relevant in dilated cardiomyopathy (DCM). Datasets stored in the GEO (Gene Expression Omnibus) database were used to assemble expanded training sets for biomarker identification analysis and classification model implementation, as well as to define independent validation datasets. The datasets consisted of samples assigned to the classes: DCM and non-DCM patients. After dataset harmonization and standardization, potential relevant features were selected through SAM and PAM analyses (Chapters 2 and 3). This process was also implemented on the original datasets independently.

The set of features detected by the SAM and PAM analyses were used to query the scientific literature with text mining software. This allowed the identification of known and potential novel associations between these genes and DCM. Using different training and validation datasets, the prediction models obtained from the aggregation of samples from different datasets outperformed the classifiers constructed on the original studies independently. Amongst the genes derived from the integrated analysis of datasets (including SAM-, PAM- and literature-derived markers), the set consisting of differentially-expressed genes without known involvement in DCM provided some of the highest classification performance using support vector machine models (Chapter 3). Moreover, its classification performance was better than that obtained with models based on known biomarkers only.

New functional characterizations of the different sets of genes were provided and interpreted in the context of a (curated) global network of protein-protein interactions implicated in DCM-related processes. Different topological and functional analyses of this network allowed the identification of additional potential biomarkers. Furthermore, this network-based analysis showed that the proteins encoded by genes suggested as potential biomarkers tend to be peripheral components of the network. This finding may corroborate that the selection of biomarkers was not biased towards well-studied (richly interconnected) genes. Additionally, it suggests that the discovered biomarkers may represent downstream response components, which reflect the effects of systemic perturbations leading to the emergence or progression of the disease.

Xu *et al.* (2008) proposed an alternative approach to discovering prognostic biomarkers from (independently-generated) microarray datasets, which is driven by gene expression data clustering for the identification of disease-specific co-expression networks. One of the key outcomes of their research was a module consisting of genes linked to tumour suppression. Although such associations have been previously established for these genes, Xu *et al.* (2008) were also able to characterize their coordinated interplay at the transcriptional regulatory level. In addition, they also moved beyond the idea of detecting gene-gene relationships through the calculation of co-expression between pairs of genes.

Using 32 microarray datasets derived from 12 tissue types (cancer types) and 23 datasets from non-cancer tissues, their integrative prediction methodology was implemented through the following tasks (Xu *et al.*, 2008): (1) A network of differential co-expression was generated, (2) the resulting network was analyzed in terms of different co-expression properties and topological attributes, and (3) network modules strongly linked to specific cancer types were predicted and characterized. In the co-expression network, a node represented a gene, and an edge linking two genes indicated that the genes were frequently co-expressed in the cancer datasets. The analysis of 'co-expression dynamics' (Xu *et al.*, 2008) was done through the calculation of 'second-order expression similarity' values between the genes across the different datasets. The co-expression between two genes, $a$ and $b$, in a single dataset is a first-order measure of expression similarity (Chapter 2). Second-order expression similarity estimates the correlation of two pairs of genes, such as ($a$ and $b$) vs. ($c$ and $d$), across multiple phenotype-specific datasets. In this way second-order similarity analysis may support the identification of co-expressed genes that may not be found by standard co-expression analysis. Using hierarchical clustering, Xu *et al.* (2008) then proceeded with the identification of second-order clusters, which enabled the discovery of phenotype-specific modules, that is modules that tend to be more active in a specific type of cancer. Thus, these modules provided a collection of biomarkers and a description of their coordinated transcriptional behaviour in specific types of cancer.

## 8.4 Data integration at the model level

As discussed in Chapter 3, support vector machine models offer many advantages for the implementation of powerful classification systems using different types of data. This makes single (or ensembles) of support vector machine models particularly suitable for heterogeneous data integration. Furthermore, this type of model allows the combination of diverse features as part of the model's intrinsic feature representation schemes. To put it another way, high-dimensional features originating from different data sources and encoded in different formats can be simultaneously combined into a single classification model without actually aggregating the original features at the input level (i.e. simple aggregation of values into a single input vector), or without the need to implement multiple models built on the different datasets independently. On the other hand, this does not imply that such traditional integration strategies cannot be implemented using support vector machine-based models.

Daemen *et al.*'s (2009) illustrated the integration of different types of 'omic' data, such as gene expression and CNV data, using a support vector machine to predict outcomes in rectal and prostate cancers. Instead of combining datasets at the input level or implementing a serial analysis of these datasets, their approach directly integrates the different features within the mathematical model of the support vector machine. The different features originating from the diverse datasets were processed 'equally' within the model, as part of the selection of the most relevant features. This integration strategy and alternative versions requires the transformation of each dataset into a 'kernel matrix' (Lanckriet *et al.*, 2004). This is a fundamental processing step as the integration of datasets is performed at this level.

Recall that the support vector machine model maps a dataset **X**, composed of $n$ samples and $m$ features, from its original input space of size $m$ into a feature space of higher-dimensionality, $q$, (with $q > m$), based on a mathematical function known as the kernel function (Chapter 3). The kernel function calculates the inner product between all pairs of samples in **X**. The resulting values are stored in the kernel matrix, which has a size of $n \times n$ values. Thus, each dataset, independently of their number of features and characteristics, can be represented by a common data representation format using the kernel matrix. This harmonized data representation within the support vector machine model is what allows the integration of the datasets in a straightforward fashion.

Daemen *et al.*'s (2009) study integrated the different datasets by computing a single, integrated kernel matrix, whose values were computed by summing up the multiple source-specific kernel matrices. Moreover, their integration approach applied a linear kernel function that generated normalized values to ensure that the different kernel matrices, representing the different types of 'omic' sources, included values measured on the same scale.

In Daemen *et al.*'s study, the integrative prediction approach was implemented using a variant of support vector machine: The 'weighted least squares support vector machine' (LS-SVM). In comparison to other versions of support vector machine models, the LS-SVM has been shown to be faster and relatively easier to implement. But more importantly, the LS-SVM is particularly suitable for unbalanced two-class prediction problems, that is applications in which the number of samples belonging to one of the classes is much larger than in the other class. The LS-SVM addresses this problem through the estimation of class-specific weights that are assigned to each sample in the dataset under analysis.

## 8.5 Multiple heterogeneous data and model integration

An example of the combination of information extracted from electronic medical records (EMR) and gene expression datasets was reported by Chen *et al.* (2008b). The EMR contained more than 1 million measurements from hundreds of laboratory tests conducted in a hospital. The goal of their investigation was to discover new biomarkers of human maturation and aging. First, the absolute count of lymphocytes was found to be strongly associated with aging. This was possible through multiple comparisons between groups of individuals on the basis of the data from their clinical laboratory tests. Several diseases, such as asthma and diabetes, were then ranked according to the levels of change of this biomarker observed in each disease. Gene expression datasets associated with these diseases were obtained from the GEO database (Chapter 9), and were analyzed to identify genes whose expression profiles were correlated with the levels of change of the clinical biomarker. Thus, this integration strategy identified sets of genes that were strongly correlated with changes in the clinical biomarker of aging, that is the absolute count of lymphocytes, including many genes known to be implicated in ageing processes.

Based on the assumption that similar phenotypes can be caused or mediated by functionally-related genes, the Web-based platform Endeavour (Aerts *et al.*, 2006; Tranchevent *et al.*, 2008) enables the prioritization of a list of candidate genes based on their functional similarity with a list of (training) genes known to be implicated in the disease investigated. Similarity is estimated between candidate and training genes using information encoded in several public information repositories, such as ontology-based functional annotations, protein-protein interactions and gene expression datasets. The calculation of similarity scores between candidate and training genes is done through different metrics adapted to the different types of data associated with a gene and stored in the databases, for example ontology-derived terms, continuous numerical data and sequences. This integrative approach to gene prioritization can be seen as a powerful exploratory engine for biomarker discovery and selection, which can guide subsequent (more focused) discovery phases. Endeavour automatically generates gene rankings based on the functional similarity scores estimated in the context of each information source. Based on these repository-specific scores, Endeavour also generates a global prioritization of genes based on an integrated score, that is the fusion of the ranking scores derived from the different sources. A key feature of Endeavour is the application of order statistics to integrate the different source-specific ranking scores. This methodology allows the combination of the different scores computed for each gene even in the face of missing values. This is particularly relevant as it contributes to the reduction of potential bias toward the most studied genes.

In some applications the main inputs to the biomarker discovery process may comprise multiple independent datasets describing the same type of 'omic' information (e.g. microarrays), but which can be associated with different, yet interrelated, phenotypes (e.g. aetiologies, disease subtypes). Moreover, it is possible that such datasets have been measured in different types of tissues or organs. Predictive integration strategies in this type of scenario are important because they may reveal common molecular mechanisms, biomarkers and potential treatment target across different (interdependent or interrelated) pathologies. This could enable researchers to answer questions such as: Are there common biomarkers or molecular processes that can predict the occurrence or progression of heart failure of different aetiology, such as cardiomyopathies and ischaemic heart disease? Are there shared mechanisms, biomarkers or therapeutic targets relevant to metastasis across different types (or sub-types) of cancers?

Ptitsyn, Weil and Thamm (2008) addressed the second question by analyzing public microarray data obtained from colorectal and breast cancer samples. Their results suggested that regardless of the tissue of origin or cancer type, metastatic tumours exhibit significant perturbations in energy metabolism, cell adhesion, antigen presentation and cell cycle regulatory pathways. In addition, their study showed that oxidative phosphorylation is significantly diminished in metastases in relation to solid tumours.

Ptitsyn, Weil and Thamm's investigation (2008) began by selecting a relative large set of potential differentially-expressed genes in the different datasets. This was followed by the detection of biological pathways and processes overrepresented in this gene list. This was done with different public and commercial biological information repositories and software tools (Chapter 9). This analysis identified 19 pathway maps and 6 clusters of functionally-related genes differentially perturbed in metastatic vs. non-metastatic tumours. Amongst these processes and pathways, oxidative phosphorylation and different types of cellular and extracellular remodelling were identified. Different visualization displays of pathways and processes conserved in metastatic tumours in different cancer types were provided. This aided in the identification of potential therapeutic targets for the prevention of metastasis. The authors suggested, for instance, that the therapeutic targeting of glycolytic pathways may represent a useful preventive strategy due to the wide conservation of perturbations in bioenergetic pathways observed in different metastatic tumours.

The problem of discovering biomarkers shared by different diseases has also been recently investigated by Dudley and Butte (2009). In contrast to Ptitsyn, Weil and Thamm's (2008) study, Dudley and Butte (2009) aimed to identify disease-specific protein biomarkers, which can be used to improve the sensitivity of prediction models. Their methodology comprised the integration of protein and gene expression data from biofluids across different diseases using a network-based approach. Blood plasma and urine biomarker networks were independently assembled. The former displayed relationships between gene expression-based profiles and blood plasma proteins in 136 diseases. The urine biomarker network encoded information associating gene expression-based profiles with proteins detectable in urine in 127 diseases. The gene expression datasets for the different diseases were obtained from the GEO. The PubMed records accompanying each GEO dataset were analyzed to extract MeSH terms, which were then used to extract disease concepts from the Unified Language System (UMLS) (Bodenreider, 2004). This process, together with text mining techniques, allowed each GEO dataset to be automatically annotated to a disease and to the tissue or biological substance of origin, and to extract control samples where available. The latter is particularly important as only microarray datasets including both disease and control samples were considered for subsequent analyses. The automated harmonization of (platform-specific) microarray probe identifiers and Entrez GeneID identifiers was carried out with the AILUN system (Chen, Li and Butte, 2007). The collection of human blood plasma proteomes was derived from the HUPO Plasma Proteome Project (PPP) (Omenn *et al.*, 2005). The proteomic data from urine was acquired from the Max-Planck Unified (MAPU) Proteome Database (Zhang *et al.*, 2007b) and the Urinary Exosome database (Pisitkun, Shen and Knepper, 2004).

In the resulting networks, nodes represented genes, proteins and diseases. An edge between a gene (protein) and a disease indicated that the gene (protein) was differentially expressed in the disease. Dudley and Butte (2009) found that more than 80% of putative protein biomarkers can be associated with multiple disease conditions, that is disease-specific biomarkers may only be found in a small subset of the proteomes measured in biofluids. This finding indicates that the search for disease-specific protein biomarkers in biofluids can be more complex than has been anticipated.

The large-scale integration of gene expression and SNPs data from genome-wide association analysis (Chapter 4) has also been recently investigated to guide the identification and validation of new disease biomarkers. For instance, Chen *et al.* (2008b) showed that highly differentially expressed genes tend to encode DNA variants associated with disease. This type of investigation can be useful to help researchers to prioritize candidate SNPs in genome-wide association studies. Chen *et al.* (2008b) demonstrated this application by ranking SNPs related to (type 1 and type 2) diabetes. Moreover, based on the analysis of highly differentially expressed genes, Chen *et al.* (2008b) were able to 're-discover' several known gene-disease associations and to distinguish true positive from false positive disease markers in different genome-wide association studies and diseases. In this investigation the microarray datasets were obtained from the GEO. The SNPs-disease associations were obtained from the Genetic Association Database (Becker *et al.*, 2004) and the Human Gene Mutation Database (Stenson *et al.*, 2003). The relationship between SNPs and gene expression was detected by means of the 'differential expression ratio' (DER). For a given gene variant, the DER was defined as the number of GEO datasets in which the gene containing the variant was found to be differentially expressed, divided by the number of GEO datasets that included the gene.

## 8.6 Serial integration of source and models

The combination of clinical and protein expression data based on a serial integration scheme (Figure 8.1e) was implemented to predict early mortality of patients undergoing kidney dialysis (Knickerbocker *et al.*, 2007). This idea was motivated by the lack of single traditional clinical and molecular markers capable of accurately predicting the outcome, that is death within the first 15 weeks of treatment initiation. These predictions are important as they can guide the definition of priorities for kidney transplantation and adaptation of dialysis treatment, such as treatment frequency or dosing. Using clinical data only, the prediction model proposed by Knickerbocker *et al.* (2007) begins with the classification of patients into two major classes: low and medium-high risk. Those patients assigned to the medium-high risk class represented the inputs to a risk stratification module based on protein expression data only. Their proof-of-concept study included data from 468 patients, who were assigned to the classes: survival and death after initiating dialysis. Amongst those patients, 208 died and 260 survived within 15 weeks of initiating the treatment. The protein expression data were acquired from 14 cytokines and other blood proteins with suspected roles in kidney disease. Eleven demographical and physiological variables, such as gender and blood pressure, represented the clinical dataset. As part of an exploratory phase, exhaustive search in logistic regression models was implemented to identify sub-sets of features in each dataset independently. Age, diastolic blood pressure, serum albumin and the method used for vascular access in the patient were the most relevant features observed in the clinical dataset. In the protein expression dataset, three cytokines were highlighted as potentially relevant biomarkers: angiogenin (Ang), interleukin-12 (IL-12) and vascular cell adhesion molecule-1 (VCAM-1). These subsets of features were then analyzed using other data mining techniques prior to their integration. Subtle, non-linear associations between each of these features and the outcome of interest (the log-odds of death) were investigated using mathematical polynomial functions known as 'splines'. These models and relationships provided the basis for the serial integrative prediction analysis. Indeed, the serial analysis approach was not proposed a priori, but was actually motivated by the non-linear relationships observed between the clinical and molecular biomarkers and between the biomarkers and the outcome. Thus, cytokine levels were shown to be most powerful for the assessment of patients classified by the clinical biomarkers as being at higher risk of death. When the clinical biomarkers identified a low risk patient the cytokine levels did not bring additional predictive information. This is an example of how integrative data analysis can guide the design of novel classification strategies tailored to specific patient sub-groups.

The success of an approach to the integration of heterogeneous data sources can also be assessed by its capacity to outperform single-source prediction models (e.g. standard clinical data only), to avoid over-fitting to a particular type of data (e.g. a microarray dataset is not particularly favoured in the integration strategy), and to deal with different types of data (e.g. clinical categorical data and continuous numerical gene expression data). Another important challenge in heterogeneous data integration is the detection of potential redundancies or dependencies between molecular and clinical biomarkers. Relatively more attention has been given to this problem in the context of feature selection in single-source, high-dimensional datasets (e.g. gene expression data). However, small sets of clinical variables and much larger sets of 'omic' features may be strongly correlated between them or redundant for predictive purposes. This is because different 'omic' features may directly influence clinical variables or vice versa, and because unknown (or unobserved) systems-level mechanisms may actually influence many of these features in common.

Boulesteix, Porzelius and Daumer (2008) offered an example of an approach designed to address these challenges based on a serial, multi-step integration of data sources and machine learning models. They aimed to demonstrate the advantages of or the conditions under which microarray data can contribute additional predictive power to applications involving traditional clinical features. Basically, their approach consisted of multiple steps of data dimensionality reduction and classification (Chapters 3 and 6) applied to microarray and clinical data. Moreover, one of the data sources was used to 'pre-validate' the classification model built on the other dataset. Feature dimensionality reduction was first performed on a (learning) gene expression dataset. The resulting features were aggregated with a (learning) dataset of clinical features to build a classifier, $M$. In a second phase, a test gene expression dataset was 're-encoded' using the results (weights) of the transformation procedure performed on the learning gene expression dataset. This transformed (gene expression) testing dataset was fused, through input vector aggregation, with the test dataset of clinical data. The classifier $M$ was then applied to make predictions using the resulting integrated test dataset. The partial least squares technique (Chapter 6) was applied for dimensionality reduction and random forests for classification (Chapter 3).

A pioneering example of the discovery of sub-network biomarkers for prognostic applications was reported by Chuang *et al.* (2007). They used network information not only to build the prediction models, but also to represent the outputs generated by these models. They demonstrated that their approach could outperform systems based on lists of (gene) biomarkers. Moreover, their research suggested that network-based biomarkers could be more robust and reproducible than traditional methods in the classification of metastatic and non-metastatic breast tumour samples. One reason to explain this predictive power and robustness is that many important drivers of metastasis do not actually show differential expression patterns between metastatic and non-metastatic tumours. The gene expression data analyzed was obtained from two independent, previously-published investigations involving breast cancer patients (van de Vijver *et al.*, 2002; Wang *et al.*, 2005).

Chuang *et al.* (2007) assembled a protein-protein interaction network consisting of more than 12 000 proteins and more than 57 000 interactions. This information was obtained from different public datasets representing experimental and computationally predicted interactions, together with curated literature-derived interactions. The first analytical phase involved the projection of the expression values from each gene onto its corresponding protein in the network. A search algorithm was used to detect sub-networks with gene expression patterns (changes) statistically associated with the group of metastasis patients. Before doing this, two basic questions needed to be answered: how to summarize the match between a patient and a sub-network? How to score the classification potential of a sub-network on the basis of the patient-network matching? For a given sub-network, the expression values from a patient (corresponding to the proteins in the sub-network) were averaged. In this way, for each candidate sub-network, each sample (patient) was represented by a unique 'activity score'. These score values were then used to compare the two phenotypes (metastatic and non-metastatic breast tumour samples) using statistical techniques, such as the t-test or mutual information coefficients. Therefore, the resulting statistic values, such as t-statistics or mutual information coefficients, could be used to describe (and rank) the classification potential of the different sub-networks. The statistical significance of the discriminatory capacity of a sub-network can be estimated by comparing the observed classification potential scores with those obtained from sub-networks randomly generated. Under Chuang *et al.*'s (2007) approach, different standard classifiers, such as logistic regression or support vector machines, can be implemented. In this scenario, each sample can be represented by its activity scores derived from the different discriminatory sub-networks.

Another example of the combination of interactome and gene expression data was illustrated by an approach proposed by Mani *et al.* (2008). Their goal was the identification of genes implicated in the oncogenesis of B-cell lymphoma. At the core of this framework was the 'interactome dysregulation enrichment analysis' (IDEA) algorithm, which combines information extracted from networks of molecular interactions together with gene expression profiles. The molecular network information was extracted from the B-cell interactome (BCI) database (Lefebvre *et al.*, 2007). BCI includes different types of transcriptional, signalling and protein complex interactions occurring in the human B cell. The IDEA-based integrative discovery framework consisted of two phases implemented serially. First, a large collection of gene expression data from normal, tumour and experimentally-modified B-cells was used to discover relevant phenotype-specific interactions in the BCI database. The aim was to find interactions exhibiting significant gains or losses of gene expression correlation in relation to a specific phenotype. Loss-of-correlation (LoC) and gain-of-correlation (LoC) information for each pair of interactions was then used to estimate statistical enrichment of LoC/GoC in groups of interactions. This allowed, in the final step, the ranking of the genes according to the LoC/GoC enrichment observed in their interaction neighbourhoods. LoC and GoC values were estimated by calculating the mutual information (Steuer *et al.*, 2002) between pairs of genes with regard to a specific phenotype, such as a cancer class. The outputs of this procedure were sets of molecular interactions that were significantly deregulated or perturbed in a specific pathology or clinical condition. Thus, such perturbed (pathway-specific) genes might be used not only as biomarkers of disease or treatment response, but also as potential therapeutic targets.

The problem of finding unique or disease-specific biomarkers can also be explored within an integrative framework consisting of multiple analysis stages implemented serially, and using a single type of 'omic' data as the initial input to this framework. An example of this approach was provided by Yang *et al.* (2008) with different gene expression datasets. Yang *et al.*'s (2008) biomarker discovery procedure began with the detection of extracellularly over-expressed genes in relation to normal cells, and in different types of tumours: prostate, breast, lung, colon, ovary and pancreas. This was followed by a filtering stage that focused on the selection of biomarkers expressed in blood (serum or plasma) and known to be involved in human cancers. The third analysis phase determined common biomarkers shared by every pair of cancer type investigated. The outcome of this phase guided the implementation of the final phase: the identification of unique (blood-borne) biomarkers for each cancer type. The over-expressed genes were obtained from the Oncomine database (Rhodes *et al.*, 2007), and the second filtering phase was implemented with the Ingenuity Pathway Analysis (IPA) system (2009). The pair-wise comparison of datasets to identify common biomarkers reported sets of 20 to 134 genes. The final analysis stage identified sets consisting of 3 to 59 genes representing potential tumour-specific biomarkers. These types of methodology and findings, and the ones reported by Ptitsyn, Weil and Thamm (2008) and Dudley and Butte (2009) (Section 8.5), may provide comprehensive and powerful insights into the search for more specific and sensitive biomarkers.

## 8.7 Component- and network-centric approaches

This categorization is suggested here to distinguish between integrative data analysis approaches on the basis of how they represent their predictions or discovery outcomes, that is biomarkers or classifications.

In the component-centric approach the main outcomes of the biomarker discovery process typically consist of a list of biomarkers, whose relationships or interactions (at any level of biological complexity) are not explicitly displayed in the prediction model or associated with the data originating from each patient at prediction time. However, it is evident that component-centric approaches can also make use of network-based information or techniques at any stage during the biomarker discovery process for different purposes. For example, networks of gene-gene, protein-protein or gene-disease associations can be used to organize or filter available knowledge, to guide the search of sets of potential biomarkers, or to generate the inputs to machine learning algorithms (Ptitsyn, Weil and Thamm, 2008; Camargo and Azuaje, 2008; Dudley and Butte, 2009).

In the network-centric approach the predictions made on a specific sample explicitly use or display information about relevant functional relationships between biomarkers. Within this category we can also include procedures in which the main outcome of the biomarker discovery process is a network or set of networks associated with specific clinical classes, conditions or patients. Examples of typical network-centric approaches are described in (Xu *et al.*, 2008; Chuang *et al.*, 2007; Mani *et al.*, 2008).

Notwithstanding the demonstrated power and success of the component-centric approach, there is a growing interest in the application of network-centric methodologies. One reason for the investigation of the latter in biomarker and drug target discovery is that key functional genes or proteins, as well as key mediators of communication, do not necessarily have to be differentially expressed in the phenotypes investigated. On the other hand, those genes showing statistically detectable differential expression do not always play key regulatory or control roles. For instance, in cancer, many genes encoding important genetic mutations may not be identified by analyzing differential gene expression. Nevertheless, these genes may have potential causative roles in the disease or can act as interconnecting components between genes, which can show differential expression patterns or more subtle functionally-relevant changes in expression. Also this may explain the robustness and reproducibility of network-based approaches in comparison with traditional gene expression bio-signatures. The latter tend to be highly variable across populations samples and independent studies. This may be, in part, explained by the possibility that many differentially-expressed genes could mainly represent downstream effectors or reactors of disease (Chuang *et al.*, 2007). In contrast, the patterns of expression changes in genes with significant influence in the emergence or progression of a disease may be more subtle than anticipated.

## 8.8 Final remarks

The integration of multiple datasets and knowledge resources has become a fundamental means to achieve more meaningful and powerful prediction and classification models. This is motivated in part by the complexity of common diseases, which can be associated with many ‘omic’ and environmental factors of small effects or a few factors of relative larger effect. On the other hand, our capacity to incorporate more features and datasets into biomarker discovery and prediction model design has significantly increased the possibilities to detect spurious relationships or decrease our capacity to reproduce potential novel findings. Although some authors have recently suggested that computational power is becoming a ‘bottleneck’ in the development of integrative data analysis for biomarker discovery (Tang *et al.*, 2009), there are probably greater fundamental concerns, such as those relating to the selection and quality assessment of the data sources. Also there is a need to incorporate expert and domain-specific knowledge to guide model construction and interpretation. Moreover, this also presents challenges from the educational and scientific culture perspectives. As part of a new generation of translational researchers, computational biologists and bioinformaticians will be required to develop a broader understanding of experimental technologies, of domain-specific assumptions and background knowledge, and of the general process of hypothesis generation and validation in the context of multiple information integration (Tang *et al.*, 2009; Chapter 10).

Other important research questions that also deserve further investigation in bioinformatics and biomarker discovery include: How could the error distribution or predictive capability of individual models be used to select problem- and data-specific integration strategies? How can data pre-processing affect or enhance integrated prediction performance? How is feature selection related to integrated prediction capability according to different data types or formats and integration strategy? How can user-driven knowledge be exploited to guide the selection of data sources and integration strategy?

Tables 8.1 and 8.2 summarize the major types of approaches and investigations reviewed in this chapter, together with recommended reading. As pointed out above, this categorization may overlap in different aspects, and an approach could be assigned to multiple categories. The main goal was to reflect the diversity of problems, applications

**Table 8.1** Relevant examples of approaches to data and model integration for biomarker discovery, based on the type of integration strategy implemented

| Approach | Typical examples and applications | Recommended reading |
|---|---|---|
| Models based on data source aggregation at the input level | Integration of different biomarkers previously discovered in independent investigations. Integration by aggregating features into a single input vector. Standard multi-variable prediction models. | (Blankenberg *et al.*, 2006; Mamtani *et al.*, 2006; Sawyers, 2008) |
| Integration of models based on homogeneous data sources | Analysis of multiple gene expression datasets for the identification of potential novel biomarker sets; combination of predictions from different models trained on the same dataset. | (Zhang *et al.*, 2007a; Hanash, Pitteri and Faca, 2008; Xu *et al.*, 2008) |
| Data integration at the model level | Different sources, including heterogeneous data types, are combined using information encoding or manipulation procedures provided by a specific classification technique, such as kernel methods. | (Lanckriet *et al.*, 2004; De Bie *et al.*, 2007; Daemen *et al.*, 2009) |
| Multiple heterogeneous data and model integration | Diverse datasets (independently generated or representing different 'omic' data types) or their corresponding prediction models are integrated in a parallel fashion. | (Aerts *et al.*, 2006; Chen *et al.*, 2008b) |
| Serial integration of source and models | Multiple analysis of different sources are implemented sequentially, the outcomes of one stage represent the inputs to the next analysis phase. | (Knickerbocker *et al.*, 2007; Mani *et al.*, 2008) |

**Table 8.2** Relevant examples of approaches to data and model integration for biomarker discovery and prediction, based on the output encoding and visualization scheme implemented

| Approach | Typical examples and applications | Recommended reading |
|---|---|---|
| Component-centric approach | The main outcome of the discovery process is a list of biomarkers, such as genes or proteins. This is obtained independently of the specific discovery strategy implemented or resources investigated, including network-based techniques. | (Aerts *et al.*, 2006; Knickerbocker *et al.*, 2007) |
| Network-centric approach | The main outcome of the biomarker discovery process are global, class- or patient-specific networks of interacting components. | (Chuang *et al.*, 2007; Mani *et al.*, 2008) |

and solutions, as well as to highlight the most relevant design and methodological features characterizing a particular approach or investigation.

This chapter also showed that the application of integrative data analysis can support our understanding of the complex relationships (or overlaps) between different diseases. This has been motivated by evidence indicating important connections, at different 'omic' levels, between several diseases previously thought to be 'dissimilar'. Some of such apparent differences may have been the product of the application of traditional techniques for disease definition and classification, such as the analysis of anatomical features or symptoms (Dudley and Butte, 2009). The identification of novel inter-disease relationships is particularly important to aid in the search of biomarkers highly specific to a single disease or clinical condition, which in some cases may represent a major requirement for assessing their potential clinical relevance. On the other hand, the identification of biomarkers shared by different disease sub-types may be useful to facilitate the discovery of novel therapeutic targets. These targets could be relevant in the prevention or delay of pathological processes shared by different disease sub-types, such as metastasis in different cancer types (Ptitsyn, Weil and Thamm, 2008).

# Guest commentary on chapter 8: Data integration: The next big hope?

**Yves Moreau**
*Katholieke Universiteit Leuven, ESAT/SCD, B-3001 Leuven-Heverlee, Belgium*

There is no doubt molecular markers are invaluable for diagnosis, prognosis, and therapy selection and follow-up in numerous pathologies. In fact, many of the classical diagnostics lab measure molecular markers. Examples include oestrogen and progesterone receptors and HER2/ErbB2 in breast cancer or the genetic markers BRCA1 and BRCA2 in familial breast cancer. For the sake of concreteness, I focus on breast cancer as one of the most active areas for complex molecular models, but it applies to many other pathologies. (An example is a statistical or machine learning model combining measurements from multiple molecular markers.)

What has changed in the past decade is the capacity to measure the genome and transcriptome genome-wide and the proteome and metabolome on a large scale. This has led to the expectation that unprecedented insight into the molecular mechanisms of complex pathologies and superior predictive models were just around the corner. The results so far have been humbling at best... For example, in breast cancer the main classification of breast tumours remains that based on the oestrogen and progesterone receptor status and the HER2/ErbB2 status. While some more complex models are taking hold, such as the 70-gene signature of van't Veer and colleagues (Cardoso *et al.*, 2008) or the 21-gene assay of ONCOTYPE-DX (Paik *et al.*, 2004), they have certainly not revolutionized our understanding of breast cancer and only improved incrementally its management. It could even be reasonably argued that complex molecular models do not outperform more classical clinical models (Edén *et al.*, 2004).

Part of this disappointing state of affairs is explained by the fact that the process of translating such molecular models into effective clinical tools is a long and challenging process. For example, the relevance of the oestrogen and progesterone status to breast cancer goes back to the early 1970s and that of HER2/ErbB2 to the mid-1980s, which demonstrates the meagre record of classical biomarkers strategies, or rather the difficulty of the task, so that there is certainly room for improvement. Complex models are hard to validate and in fact require large clinical trials, such as the MINDACT trial for breast cancer (Cardoso *et al.*, 2008). Tough economic realities will thus limit the number of models that will make it into clinical practice. However, we can expect that models will slowly but steadily hobble through the development pipeline and that an increasing number of clinically relevant models will become available over time.

There is, however, a number of challenges specific to complex molecular models. First, a truly clinically relevant outcome needs to be addressed. When building molecular models, we are faced with the complexities of the molecular data and the limitations in the annotation of outcomes. Complex models are 'sample hungry' and access to many samples may be a real challenge forcing us to predict surrogate outcomes (for example, response to chemotherapy instead of the more relevant ten-year survival). There is a risk that the model will not be effective when we try to redesign it for the outcome of actual interest. Second, a basic assumption of most statistical and machine learning predictive modelling is that the data to which the model is applied are drawn from the same distribution as the data from which the model was designed. This assumption is essentially self-evident except it is almost never true in clinical practice! Differences in sample handling from lab to lab and patient populations across clinics almost guarantee that this basic assumption cannot hold. The problem of dealing with such messy data is challenging and has been given only limited attention, yet it should be a priority. Simpler models based on only a few well-characterized variables and put together and refined over time by clinical experts have robustness implicitly built in. Third, another key reason why models based on 'omics' data have found it challenging to outperform models based on clinical data (which will actually often contain a number of key molecular markers) is that such clinical data are often a close reflection of the phenotype. While 'omics' data have been hailed for their potential to unravel the molecular cascades underlying a phenotype, it may not necessarily translate into superior predictive power. For example, why try to predict the invasiveness of a breast tumour from transcript levels when you can directly measure whether it has invaded the satellite lymph nodes next to the breast? By contrast, a strong example of molecular model is for the staging of breast tumours. While staging based on histopathology is difficult for many tumours, leveraging proliferation as the key biological process for staging has led to a significant improvement of breast tumour management (Loi *et al.*, 2007).

Two major goals that can be tackled with molecular models are discovery and prediction. In discovery mode, the goal is rather to identify molecular cascades and key biomolecules implicated in the disease process. We can view this also as a gene-centric point of view. In prediction mode, the goal is rather to predict a clinical outcome with robust accuracy (not necessarily the same as highest accuracy). We can view this as a patient-centric point of view. These two activities are not strictly separate and flow into each other. In practice, we will probably start from fairly complex discovery models and slowly extract the key aspects of those models to build simple robust predictive models. This may not mean that the predictive models are reduced to only a few markers, but probably to only a few sets of interrelated markers (often called modules) because modules are likely to provide a more robust measurement of the state of a cascade than the best performing single marker.

This chapter addresses integrative strategies for modelling across multiple types of 'omics' data (genotyping, copy number, epigenetics, binding, transcriptomics, proteomics, and metabolomics). With the increasing availability of 'omics' data, we may be tempted to 'throw everything but the kitchen sink' at these difficult problems.

Can we expect any true improvement from complex data integration or is it rather a matter of raising the stakes when playing a losing hand? An unfavourable aspect of the problem is that the collection of multiple data types in a clinical context is likely to remain extremely challenging for a long while. Even if the cost of the underlying technology decreases rapidly, sample handling across multiple platforms will remain a significant bottleneck. A possible breakthrough would be if next-generation sequencing made it possible to produce multiple types of data from a single sample (genotyping, copy number, epigenetics, binding, transcriptomics) in a highly automated fashion. Except if such a breakthrough occurs, complex integrative models are more likely to be useful for discovery than for prediction.

On a more positive note, in integrative models there are actually two key aspects that provide a new angle of attack to the problem: (1) availability of multiple types of hopefully complementary data that give us a more comprehensive view of the biological phenomenon, and (2) a network approach to the problem. First, an important aspect of learning from multiple types of data for gene-centric discovery is that it can be carried out using different biological samples for different data types (in patient-centric prediction, this would probably make little sense). This provides much additional statistical power for discovery. Second, the network and module view is likely to help to significantly reduce problems with false positives. While the scale of 'omic' measurements make it difficult to distinguish relevant markers from irrelevant ones, significant modules of highly interconnected genes are likely to be truly relevant. Moreover, we can often associate molecular roles to such modules, which add an important semantic level to the model and makes them more likely to be accepted in clinical practice. It is one thing to identify a loose set of genes predictive of tumour prognosis and another to state that this set represents the activity of the key process of proliferation. This has already been the aim of cluster analysis, with mitigated results, but it is likely to be more successful via network analysis because network sparseness makes for tighter patterns.

Finally, an intriguing observation about the serial integration presented in this chapter is that it is somewhat reminiscent of differential diagnosis, which is central to medicine. Essentially, differential diagnosis is a decision tree where each node consists of one or more clinical observations or tests and leads to a final refined diagnosis. It applies also to therapy decisions. Little of that is currently reflected in our predictive models. Relevant clinical models for multiple data types may eventually very much resemble such procedures. For example in breast cancer: To which of the basic classes does a tumour belong? For a tumour of this particular class, does it carry this specific mutation? For a tumour of this class with the mutation, does it show, for example, a strong immune response? In this case, select this particular treatment. Each question is best addressed by looking at a different type of data, which makes integrative analysis essential. Such models would fit tightly with clinical practice where physicians need to make and communicate informed decisions to patients whose health is at stake.

This chapter discusses current strategies for integrative analysis from this discovery to prediction. We can expect a long road before such sophisticated models truly affect clinical practice, but by leveraging complex data in a structured way we can ultimately expect an important clinical impact. Even if the resulting clinical models eventually appear a lot simpler, complex strategies are likely to speed up their design. Even if omniscient 'omics' models may still be a long way off, we must recognize that any robust significant improvement of the clinical management of a pathology is a major medical advance that will affect the lives of thousands or millions of individuals and we should aim for that.

## References

Cardoso, F., Van't Veer, L., Rutgers, E. *et al.* (2008) Clinical application of the 70-gene profile: the MINDACT trial. *Journal of Clinical Oncology*, **26**, 729–735.

Edén, P., Ritz, C., Rose, C. *et al.* (2004) "Good Old" clinical markers have similar power in breast cancer prognosis as microarray gene expression profilers. *European Journal of Cancer*, **40**, 1837–1841.

Loi, S., Haibe-Kains, B., Desmedt, C. *et al.* (2007) Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade. *Journal of Clinical Oncology*, **25**, 1239–1246.

Paik, S., Shak, S., Tang, G. *et al.* (2004) A multigene assay to predict recurrence of tamoxifen-treated, node-negative breast cancer. *The New England Journal of Medicine*, **351**, 2817–2826.

# 9 Information resources and software tools for biomarker discovery

This chapter will discuss key information resources, software tools, data exchange approaches, analysis workflow platforms and bioinformatic infrastructures for supporting biomarker discovery research. Different open- and closed-source systems relevant to biomarker discovery will be introduced. Other key topics to be discussed are: knowledge bases for guiding the search and evaluation of new biomarkers, and extensible and integrative software systems for assisting different tasks in biomarker discovery research. It will also present examples of relevant investigations, solutions and applications in biomarker discovery.

## 9.1 Biomarker discovery frameworks: key software and information resources

The great diversity of molecular and clinical information resources not only demands more advanced data analysis techniques, but also flexible infrastructures to share and interconnect these resources and tools. This requires the design of efficient, more reliable and extensible software, as well as mechanisms to deploy both human- and machine-readable information resources and analysis services. Such infrastructures also require global, national and regional initiatives to promote human collaboration through sustainable bio-computing platforms and innovation networks. Bioinformaticians and computational scientists have responsibilities that go beyond the implementation and/or operation of data acquisition, tracking, storage and statistical analysis systems. They are also required to contribute to a variety of research support and knowledge discovery

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

tasks. This involves, for example, new computational tools for predicting new biomarkers using seemingly unrelated information resources, simulation systems to predict responses to external interventions, and the development of comprehensive well-organized knowledge bases to aid in drug target and biomarker discovery.

Despite recent progress, many biologists and clinicians currently do not have full access to key resources and tools needed to implement integrative, quantitative and predictive analyses of multiple types of information. Applications such as the combination of different datasets, including those originating from different 'omic' resources, to establish functional or diagnostic associations may require specialized training in data mining and software development. This chapter will discuss current and emerging efforts to facilitate access to information, software tools and other analytical resources relevant to the discovery and evaluation of biomarkers. Many of these efforts have been inspired or guided at some level by different principles of openness and collaboration.

The major components of a 'translational bioinformatics' infrastructure, also known as cyberinfrastructure (Stein, 2008), are data resources, computational tools and applications that are interconnected via communication platforms, services and protocols. Figure 9.1 illustrates a hypothetical example of a translational bioinformatics infrastructure, which facilitates data access and analysis, knowledge generation and exchange, and research cooperation. Examples of data resources, tools, and collaborative initiative will be presented in subsequent sections. Note that the access to local and external resources and systems should be as transparent and flexible as possible to the researcher. Independently of the sophistication and diversity of resources, as well as the collaboration levels, these components rest on and are driven by the existence of a cross-disciplinary human force. This requires not only human resources to install and maintain resources, but also the active participation of physical and computational scientists and engineers in the development or adaptation of new tools and applications. The costs of deploying, adapting and sustaining infrastructures can be reduced by implementing

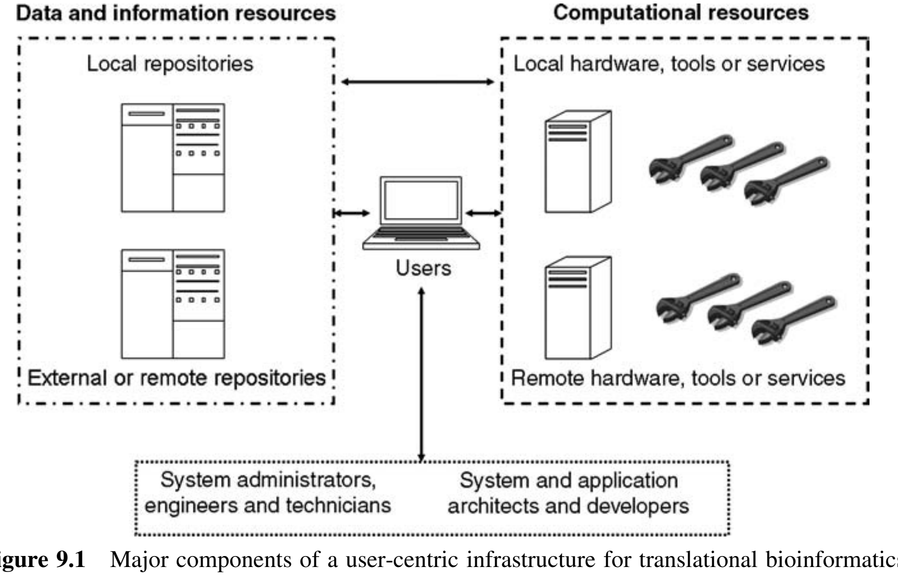


**Figure 9.1** Major components of a user-centric infrastructure for translational bioinformatics research

leading-edge technologies for sharing and optimizing informatic resources, such as computing grids and open-source software. Nevertheless, one cannot overstate the importance of fostering the formation of well-trained and coordinated research teams with the ability to work in both top-down and bottom-up management scenarios, as well as in inter-disciplinary and inter-institutional projects.

The communication infrastructure required to link these resources is more than physical devices, protocols and network connectivity. This also includes layered informational structures and platforms that implement the syntactic and semantic communication between the above resources (Stein, 2008). Syntactic integration involves the use of common data formats for different types of data and applications. Semantic integration refers to standards and technologies that allow machines to interoperate and 'understand' the context and meaning of the information. Such a semantic understanding would allow, for example, the unambiguous application of biomedical terms or concepts with multiple possible interpretations across disciplines.

## 9.2 Integrating and sharing resources: databases and tools

Tables 9.1–9.4 offer examples of representative public data resources derived from research in genomic variation, gene expression, proteomics and interactomics. Table 9.5 includes examples of key initiatives to support a more intelligent exchange, sharing and integration of 'omic' data and information relevant to biomarker discovery research. Such standards, recommendations and checklists are already reflected in most representative databases, and will continue driving future advances in resource integration or interoperability. Table 9.6 presents additional examples of information resources designed for specific disease domains. Table 9.7 highlights relevant examples of projects and tools based on open-source software. These examples highlight a diversity of efforts, technologies and purposes in supporting both basic and translational biomedical

**Table 9.1** Examples of public genomic variation databases

| Source | Data type | Access |
|---|---|---|
| COSMIC | Somatic mutations in cancer | www.sanger.ac.uk/genetics/CGP/cosmic/ |
| dbSNP | SNPs, short deletions | www.ncbi.nlm.nih.gov/SNP/ |
| GeMDBJ | Genome Medicine Database of Japan, SNPs | gemdbj.nibio.go.jp |
| HapMap | SNPs | snp.cshl.org |
| Pan Asian SNP Database | SNPs | pasnp.kobic.re.kr/ |
| Progenetix | CGH, SNPs | www.progenetix.net |
| Recurrent Chromosome Aberrations in Cancer | Chromosome aberrations | cgap.nci.nih.gov/Chromosomes |
| SKY/M-FISH and CGH Database | CGH | www.ncbi.nlm.nih.gov/sky/ |
| SNP500Cancer | SNPs | snp500cancer.nci.nih.gov/ |
| The CNV Project Data | Copy number variation | www.sanger.ac.uk/humgen/cnv/ |

CGH: Comparative genomic hybridization. SNPs: Single nucleotide polymorphisms.

**Table 9.2** Examples of public gene expression data resources

| Source | Description | Access |
|---|---|---|
| ArrayExpress | Expression profiles from curated experiments | www.ebi.ac.uk/microarray-as/ae |
| CIBEX | Search and browsing of different datasets | cibex.nig.ac.jp/ |
| GEO | Gene Expression Omnibus, curated data browsing, query and retrieval | www.ncbi.nlm.nih.gov/geo |
| GeMDBJ | Genome Medicine Database of Japan, Affymetrix expression data | gemdbj.nibio.go.jp |
| microRNA.org | microRNA targets and expression data | www.microrna.org |
| Oncomine | Cancer gene expression profiling and analysis tools | www.oncomine.org |
| SMD | Stanford Microarray Database, multiple-organisms data query and analysis tools | genome-www5.stanford.edu |

research. However, there are still significant gaps to be addressed for enabling large-scale, integrative access and analysis of disease-specific, clinical and molecular information sources (Mathew *et al.*, 2007).

Recent research (Stein, 2008) has indicated that the bioinformatics infrastructure required to support biological and biomedical research, in general, has advanced in terms of efforts for data and information sharing. However, integrative bio-computing infrastructure of tools and applications, as well as greater syntactic and semantic interconnectivity across them and information sources, are still in their infancy and will require significant investments of technological and human resources. For example, despite the

**Table 9.3** Examples of public proteomic data resources

| Source | Description | Access |
|---|---|---|
| PRIDE | Protein and peptide identifications | www.ebi.ac.uk/pride |
| OPD | Open Proteomics Database, mass spectrometry-based proteomics data | bioinformatics.icmb.utexas.edu/OPD |
| HPPP | Plasma Proteome Project, proteomic data from human serum and plasma | www.hupo.org/research/hppp |
| GeMDBJ | Genome Medicine Database of Japan, proteomics data in cancer research | gemdbj.nibio.go.jp |
| GPM | The Global Proteome Machine, tandem mass spectrometry data from different organisms | www.thegpm.org |
| HPR | Human Protein Atlas, expression and localization of proteins in human normal and cancer cells | www.proteinatlas.org |
| PeptideAtlas | Multi-organism compendium of proteomic data | www.peptideatlas.org |

**Table 9.4** Examples of public molecular network databases in humans

| Source | Description | Access |
|---|---|---|
| BioCarta | Search and visualization of signalling pathways | www.biocarta.com |
| BioGRID | Curated datasets of physical and genetic interactions | www.thebiogrid.org |
| CellMap | The Cancer Cell Map, cancer-related signalling pathways | cancer.cellmap.org/cellmap |
| HPRD | The Human Protein Reference Database, annotated protein interaction information | www.hprd.org |
| KEGG | Kyoto Encyclopaedia of Genes and Genomes, metabolic pathways | www.genome.jp/kegg |
| REACTOME | Database of core pathways and reactions | www.reactome.org |
| WikiPathways | Open, community-driven platform of curated biological pathways | www.wikipathways.org |

availability of different online databases and software, the work conducted by many translational biomedical researchers on a daily basis still depends on multiple time-consuming manual or semi-automated tasks. A major reason is that researchers need to select and to query different resources (e.g. disparate databases and packages), followed

**Table 9.5** Examples of public initiatives in information standardisation, exchange and harmonization

| Project | Description | Access |
|---|---|---|
| MIBBI | Minimum Information for Biological and Biomedical Investigations, access to checklist development projects and their products, 'one-stop shop' of minimum information checklists | www.mibbi.org |
| GO | The Gene Ontology, controlled vocabulary to describe gene and gene products, several databases annotated to the GO | www.geneontology.org |
| MAGE | MicroArray and Gene Expression Data, data formats for storage and exchange | www.mged.org |
| PSI initiative | HUPO Proteomics Standards Initiative, standards for data representation in proteomics | www.psidev.info |
| BioPAX | Data exchange formats for biological pathways: metabolic pathways, molecular interactions, signalling pathways gene regulation and genetic interactions | www.biopax.org |
| CDISC | Clinical Data Interchange Standards Consortium, standards to support the acquisition, exchange, submission and archive of clinical research data and metadata | www.cdisc.org |

**Table 9.6** Examples of public, disease-specific databases.[a]

| Project | Description | Access |
|---|---|---|
| AlzGene | Genetic association studies performed on Alzheimer's disease phenotypes | www.alzforum.org |
| EPConDB | Information about genes relevant to diabetes, pancreatic development and beta cell function | www.cbil.upenn.edu/EPConDB |
| GOLD.db | Genomics of Lipid-Associated Disorders Database | gold.tugraz.at |
| NEIBank | Ocular genomics and annotated eye disease genes | neibank.nei.nih.gov |
| T1DBase | Type 1 Diabetes Database, different types of data on susceptibility and pathogenesis | t1dbase.org |
| KinMutBase | Registry of disease-causing mutations in protein kinase domains | www.uta.fi/imt/bioinfo/KinMutBase |

[a] Tables included above also offer examples of cancer-related databases.

**Table 9.7** Examples of relevant projects based on open-source software development

| Project | Description | Web site |
|---|---|---|
| Bioconductor | Statistical analysis of 'omic' data | www.bioconductor.org |
| BioMart | Database integration system for large-scale data querying | www.biomart.org |
| BioMOBY | Support for automatic discovery and interaction with different biological databases and analytical services | www.biomoby.org |
| BIRN | Distributed virtual community of shared resources to support disease-oriented research | www.nbirn.net |
| caBIG | Integrated network of standards, resources and tools supporting cancer research | cabig.nci.nih.gov |
| Cytoscape | Extensible platform for the visualization and analysis of biological networks and other 'omic' data | cytoscape.org |
| Galaxy | Web portal for searching, querying and visualizing information stored in different remote databases | g2.bx.psu.edu |
| Taverna | Graphic environment for designing and executing workflows, which allows the integration of different online software tools | taverna.sourceforge.net |
| Weka | Software platform for the implementation and evaluation of several data mining algorithms | www.cs.waikato.ac.nz/~ml/weka |

by the independent filtering and aggregation of resulting outcomes prior to an 'integrated' interpretation. Moreover, most of the existing databases, analysis systems and Web services mainly rely on 'data-centric' solutions (Stein, 2008). This means that these resources have put greater emphasis on data access applications. Moreover, many of them are not sufficiently user-friendly and deserve improvements to support a more flexible design of integrative tasks, including the visualization and integration of resources and analysis outputs.

Recent advances in the development of integrative, community-driven information and knowledge resources include annotation systems based on the 'Wiki' principles of online publication and editing of information. One example is WikiPathways (Table 9.4) that offers collections of different biological pathways in humans and other species. These pathways can be annotated and edited by the scientific community through different user-friendly, graphical tools integrated into the system.

Current and future progress in the integration of multiple, heterogeneous information resources will be supported by Web services and Semantic Web solutions (Stein, 2008). Web services exploit standards for describing the capability of applications and for invoking them when needed. Aside from such standards (e.g. Web Services Descriptor Language, WSDL, and the Simple Object Access Protocol, SOAP), there are collections of open-source software libraries and applications that allow developers to create, search and run Web services in different applications. Two of the best known Web service development systems in bioinformatics are the Globus toolkit (Sotomayor and Childers, 2005) and BioMOBY (Wilkinson and Links, 2002). Unlike Web services, Semantic Web applications do not define a strong distinction between data and the procedures applied to these data (Stein, 2008). Rather, they encode all resources and applications as pieces of information, whose relationships are explicitly defined with ontologies and exploited by 'reasoning engines' or 'reasoners'. However, the incorporation of Semantic Web applications into practical or routine applications in bioinformatics is still the subject of intensive investigation and discussion.

A successful story in the application of advanced software development, together with the latest technology in knowledge management and Web services, is the Taverna project (Oinn *et al.*, 2004; Hull *et al.*, 2006). It allows researchers to design sequences of diverse bioinformatic analyses using both public and proprietary resources under a graphical and interactive environment. The Taverna system implements bioinformatic workflows: The automated search, calling, integration and execution of bioinformatic applications on user-selected data. This requires from the user a minimum of knowledge of the operation of the resources, such as statistical analysis techniques and their location. This system is a product of the myGrid project, which involves the cooperation of different universities and research centres in the UK. Taverna can be run on any computer and operating system with Java and an Internet connection, and the user does not have to install additional applications or databases.

The BioMart system (Fernández-Suárez and Birney, 2008) is a Web-based platform that supports the user-driven integration of online databases. Based on a list of databases pre-selected by BioMart, such as UniProt and Reactome, a user can select the primary resource to be queried together with specific search filters and criteria. The outcomes of this search can then represent the input to a secondary, domain-specific application, which is also selected by the user. For example, a user may integrate a pathway database search with a search for proteins constrained by GO terms or other functional attributes. BioMart plugins have also been incorporated into other software, such as Taverna.

## 9.3 Data mining tools and platforms

Most existing software tools cannot be easily integrated with other solutions, or are difficult to adapt or extend. The Cytoscape (Shannon *et al.*, 2003) project is an example of how to overcome the isolationist approach to software development and sharing. It offers a core platform for visualizing and analyzing different types of biological networks. Cytoscape is built on software development principles and methods that allow the incremental, open incorporation of new tools, software functionality or analytical techniques. This means that Cytoscape is based on a 'plug-in' architecture. The Cytoscape core system offers a diverse tool set for graphical visualization, manipulation, editing and topological analysis of different types of biological networks. It allows the importation and exportation of network data files in different formats, and can be used as an interface to different integrative data analysis tasks, including those requiring gene expression data, GO-based annotation analysis and text mining. Plugins for the quantitative simulation of biochemical pathway models and network clustering are also available. One of the most powerful functionality features offered by Cytoscape is the interactive integration of biological networks and gene expression data, which allows the user to load data, set visual properties of nodes and connections, and analyze functional and structural patterns (Cline *et al.*, 2007).

The Bioconductor project (Gentleman *et al.*, 2004) offers an open-source platform for the analysis and management of different types of 'omic' data. Bioconductor is primarily based on the R programming language. Therefore, users need to install R to allow the installation of Bioconductor and its default set of software packages. Bioconductor offers several software packages that operate as add-on modules under R. Statistical analysis packages for DNA sequence, microarray, SAGE and SNPs data are available (Gentleman *et al.*, 2005). Some data integration systems, such as BioMart, offer options to establish direct interactions between Bioconductor and different online databases (Fernández-Suárez and Birney, 2008).

The availability of software tools to assist in the analysis and interpretation of genome-wide association studies is gradually expanding (Buckingham, 2008). Most of the existing public or open-source tools tend to focus on SNPs data, but new solutions tailored to the analysis of copy-number variations will become increasingly available. The development of the next generation of tools face different challenges ranging from the massive sizes of these datasets, computing and graphic-processing power constraints, and the relative lack of computing-efficient statistical tests to filter spurious associations. Moreover, many of the existing tools may present significant usage barriers to those researchers who do not have a strong background in statistical analysis or bioinformatics. However, recent advances are addressing these and other requirements, such as the development of more interactive, visualization-driven software packages. Goldsurfer2 (Pettersson *et al.*, 2008) is a Java-based interactive and user-friendly graphical tool that can support different analytical steps in genome-wide association studies, including quality control and statistical analysis. This system can be applied to datasets including hundreds of thousands of SNPs and thousands of samples (Pettersson *et al.*, 2008).

Future advances in this area will comprise integrative data analysis approaches to the detection of biomarkers and functional pathways based on genomic variation, gene expression and protein network data. Moreover, these approaches will be designed to operate under open-source, extensible software platforms, such as the Cytoscape system (Buckingham, 2008).

Other examples of widely applied tools that cover different steps of 'omic' data analysis are the Gene Expression Profile Analysis Suite (GEPAS) (Montaner *et al.*, 2006), Weka (Frank *et al.*, 2004) and the UCSC Cancer Genomics Browser (Zhu *et al.*, 2009). GEPAS focuses on the analysis of microarray data and their interpretation with different distributed biological information resources. GEPAS enables a Web-based, automated implementation of different analytical tasks: ranging from data pre-processing, through clustering and supervised classification, to functional annotation based on the GO and biological pathway databases. It offers different user-friendly options for finding differentially expressed genes using several statistical analysis techniques, feature filtering based on feature-class or feature-survival times correlations, and methods for the combined analysis of gene expression and genomic copy number variation data.

Weka is an open-source, machine learning workbench (Witten and Frank, 2005) that has been applied to different biomedical research domains, including several biomarker discovery applications (Frank *et al.*, 2004; Azuaje, 2006). Weka provides not only a comprehensive collection of data pre-processing and supervised classification algorithms, but also different interfaces: the Explorer, the Knowledge Flow and the Experimenter, which assist the researcher in the development and evaluation of different types of applications. The Knowledge Flow, for example, offers a graphical interface based on a process-oriented design approach, in which the different data mining components can be selected and combined in an interactive workflow of components and information.

The UCSC Cancer Genomics Browser (Zhu *et al.*, 2009) enables the integrative visualization and analysis of different types of genomic and clinical data. For instance, users can create different genome-wide graphical displays and statistical analyses of DNA variation and gene expression data. This can be done together with an interactive selection of clinical features from different samples and disease types. Visual displays, such as heatmaps, can be zoomed and linked to the UCSC Genome Browser (Karolchik *et al.*, 2008) to retrieve additional information. The UCSC Cancer Genomics Browser is not limited to cancer-related data. Moreover, the system can be used as a public Web-based browser or as a locally-installed application.

A novel approach to supporting 'omic' data mining of potential disease biomarkers is offered by the Endeavour system (Tranchevent *et al.*, 2008), which is based on the idea of prioritizing genes that are implicated in specific biological processes or diseases. A typical application would involve a training dataset of genes known to be associated with a specific disease, which is used to automatically build different prediction models based on different 'omic' data sources in different organisms. The resulting models are applied to a user-defined testing or query dataset. Endeavour automatically ranks the query genes in relation to the training dataset using different quantitative prioritization scores. This system incorporates more than 50 public data resources to build the prediction models, and has supported biomarker discovery research in obesity and type II diabetes, amongst other areas (Tranchevent *et al.*, 2008). Chapter 8 presented a more detailed discussion of the Endeavour system.

## 9.4 Specialized information and knowledge resources

The availability of a diverse range of tools, data resources and applications presents users with another big challenge: How to keep track of existing solutions? How to compare and select the most appropriate application-specific options? Resource categorization and integration frameworks represent a valid approach to tackling this problem. One such system is the iTools framework (Dinov *et al.*, 2008), which provides a system for the cataloguing, classification and integration of different computational resources across different application domains and bio-computing infrastructures.

This type of integrated bioinformatic frameworks may be accessed by both humans and computing systems to search, compare, display and assess collections of resources. Moreover, these systems may facilitate a more efficient communication between developers, scientists and policy makers, as well as innovative ways to support accountability and evaluation of resources. In the case of iTools, an important feature is the capability to allow users to populate and manage the content of its integrated resource environment (Dinov *et al.*, 2008). The iTools framework also offers support for the integration of software plugins, including Web-crawlers and browsers.

The integration of bioinformatic resources is also constrained by the lack of structured information and knowledge across application domains and disciplines. As discussed above, Semantic Web approaches may contribute to a more intelligent and user-friendly navigation, management and interpretation of information in the systems biology era. The Semantic Web is based on the idea of exploiting common formats to support resource integration at different information processing levels. For example, different efforts are currently under way to offer biomedical data in well-structured formats based on ontologies, as well as prototypes of decision support systems that explore new functional features unavailable in the current Web (Ruttenberg *et al.*, 2007). However, the development of the Semantic Web tailored to translational biomedical research will continue to be constrained by the lack of semantically annotated information resources (including scientific publishing), the limited scalability of existing Semantic Web-compliant applications, and the need to develop methodologies to assess the origin and generation methods of information and knowledge resources, that is their provenance (Ruttenberg *et al.*, 2007). On the plus side, concrete applications in different biomedical research areas are being developed (Ruttenberg *et al.*, 2007), and advances with regard to patient identity processing, system security and networking infrastructures will continue to grow in the near future (Winter *et al.*, 2007).

## 9.5 Integrative infrastructure initiatives and inter-institutional programmes

Two important examples of large-scale cooperative efforts for the development of integrative bioinformatics infrastructures are the Biomedical Informatics Research Network (BIRN) and the Cancer Biomedical Informatics Grid (caBIG) in the US. These projects use Web service technologies to interconnect multi-disciplinary, geographically-distributed teams and bio-computing resources. The BIRN (BIRN, 2008; Keator *et al.*, 2008) is a virtual community of shared resources and collaborative tools for biomedical research. This initiative coordinates the implementation of open-source software tools and applications for data acquisition, analysis and management, which are shared under the BIRN infrastructure. BIRN also comprises a repository of freely-accessible data and metadata, with an emphasis on biomedical imaging, which supports data annotation, visualization and querying. As part of BIRNs efforts to support flexible and scalable resource integration, different biomedical ontologies and knowledge management collaborative tools have been developed. Such resources cover different aspects in neuroanatomy, behaviour and cognitive research, and experimental protocols. Any research group can connect to the BIRN bioinformatics infrastructure by installing a 'BIRN rack', which is a software installation and deployment system designed to facilitate the incorporation of new end-points or end-users.

The caBIG (2008) offers a bio-computing infrastructure that supports research collaborations in the cancer community, including researchers, physicians and patients. caBIG's main goal is to facilitate new advances for the diagnosis, treatment and prevention of cancer based on an integrative, collaborative environment of information and computing resources. This has required investments in the development of shareable software tools and platforms, ontologies and knowledge sharing standards. However, caBIG has given priority to the idea of open-source development and reuse of existing tools and resources to conceive domain-specific solutions. One of the resulting contributions of caBIG has been its support for the development of 'The Cancer Genome Atlas' (TCGA). The caBIG project has also supported the development of a bioinformatics platform for sequence and expression data analysis, the geWorkbench system. The geWorkbench (2008) uses a plugin architecture that offers different tools for data analysis and their interaction with other packages, such as the BioConductor and Cytoscape systems. The caGrid is the network of Web services deployed by caBIG.

The Early Detection Research Network (EDRN) is an initiative of the US National Cancer Institute, which aims to support the incorporation of new molecular diagnostics and biomarkers into personalized medicine of cancer patients. This network offers a Web site and different resources to inform and coordinate efforts in biomarker research (EDRN, 2008). Similarly, based on public-private partnerships, the EDRN aims to support studies, including trials, to validate biomarkers for the early detection of cancer, risk assessment and their application as surrogate endpoints.

A catalogue of recent projects and funding initiatives, which are being developed in the European Union (EU), related to bioinformatic infrastructures and biomarker research has been produced by Marcus (2008). Different databases, tools and services are being developed as part of the EU Framework Programme, which promotes collaborative research between the public and private sectors. Examples of recent advances include new resources for distributed computing, *in silico* modelling toolkits, text mining and prototypes of systems biology 'toolboxes', which tend to focus on the study of different model organisms. Significant contributions are also represented by efforts to integrate bio-computing resources and human expertise, such as the European Bioinformatics Grid and the European School of Bioinformatics (Marcus, 2008).

## 9.6 Innovation outlook: challenges and progress

Significant advances in terms of information resource integration in specific research areas have been accomplished over the past 10 years. The next transformative phase to accelerate translational research, in general, and biomarker discovery, in particular, will comprise other levels of integration. These new levels should include software tools, applications and findings across and within disciplines. There is evidence that, at least in the near future, disciplines will continue developing their information sharing, management and reasoning capabilities in relative isolation through advances in Grid computing, open-source software development and knowledge representation (Stein, 2008). New advances will also depend on the maturation of emerging technologies, such as the Semantic Web, and the further development of a culture of electronic collaboration and information sharing. The latter does not only involve data and software utilities, but also hardware and knowledge in the form of annotations and machine-readable literature collections.

As depicted in Figure 9.2, a crucial goal is to integrate different prediction models. Thus, advanced bioinformatic infrastructures should provide tools to support the implementation of four main tasks: Data exploration and selection; model design and implementation, model evaluation and selection; and validation studies. As discussed above, advanced bioinformatic infrastructures will continue to be developed as open, extensible software workbenches (Figure 9.3). Such integrated, extensible platforms may consist of a 'core system' and an 'application (plug-in) system'. For instance, a core system may comprise: (a) a set of generic software components; (b) a prediction model discovery engine; and (c) a validation and application engine. The generic software components allow fundamental tasks, such as input/output data manipulation. The prediction model discovery engine can include application-independent components to allow the selection of prediction models, their design and evaluation and output documentation. The validation and application engine should provide generic components for importing and exporting independently-evaluated models. Under such an integrated, extensible software framework, a second major system component is the application plug-in system, which can allow the incremental integration of tools, such as

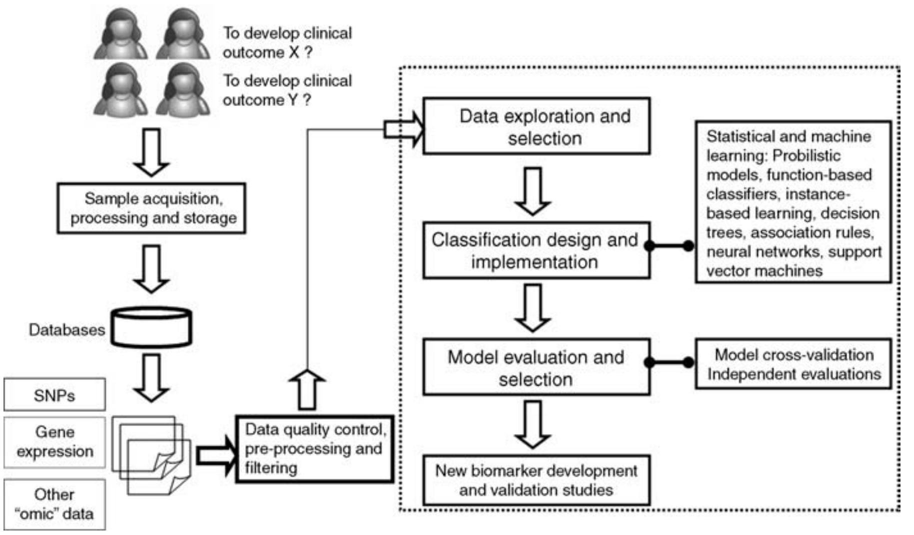


**Figure 9.2** An example of a typical biomarker discovery data analysis workflow within an integrated bioinformatics framework

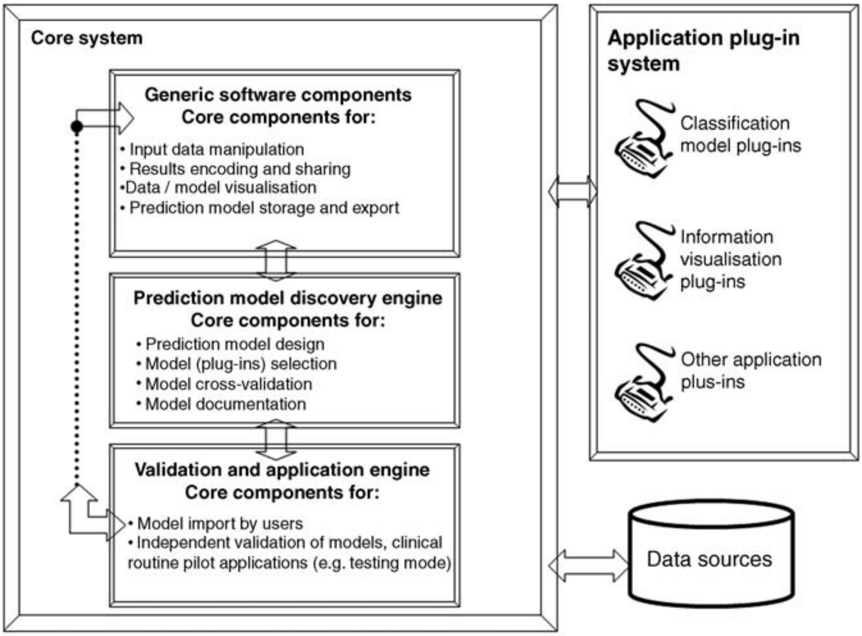


**Figure 9.3** An example of an integrated, extensible bioinformatics infrastructure: Overview of a typical software architecture

classification models and information management methods. Therefore, this system architecture can allow the independent implementation of plug-ins and their integration into the infrastructure without requiring the end-user to directly modify the existing software code.

Figure 9.4 presents a simplified view of a minimum set of resources, tools and outcomes that are required in the development of a bioinformatic infrastructure for supporting biomarker discovery and evaluation. Moreover, such building blocks are required to achieve different degrees of integration and coordinated communication at different levels. This type of framework can shape the development of different applications relevant to biomarker discovery and evaluation in different biomedical domains, including support for potential drug target identification and clinical trials.

Also under this framework, diverse resources of molecular and clinical data acquired from prospective and retrospective studies can be integrated locally and remotely. These resources are required to comply with information representation standards and formats that make them compatible with a variety of software tools and applications. Software resources that should be integrated range from data acquisition and tracking, through data mining and management packages, to knowledge annotation engines.

Another important requirement is the incorporation of workflow management systems or automated design workbenches to assist researchers (with different specialization backgrounds) in information-driven prediction model selection, implementation and evaluation. Furthermore, integration at the prediction and hypothesis generation levels will continue driving innovative advances. This goes beyond the physical or virtual integration of bioinformatic resources. As shown in previous chapters, this also consists of the combination of different information displays, predictions and interpretations

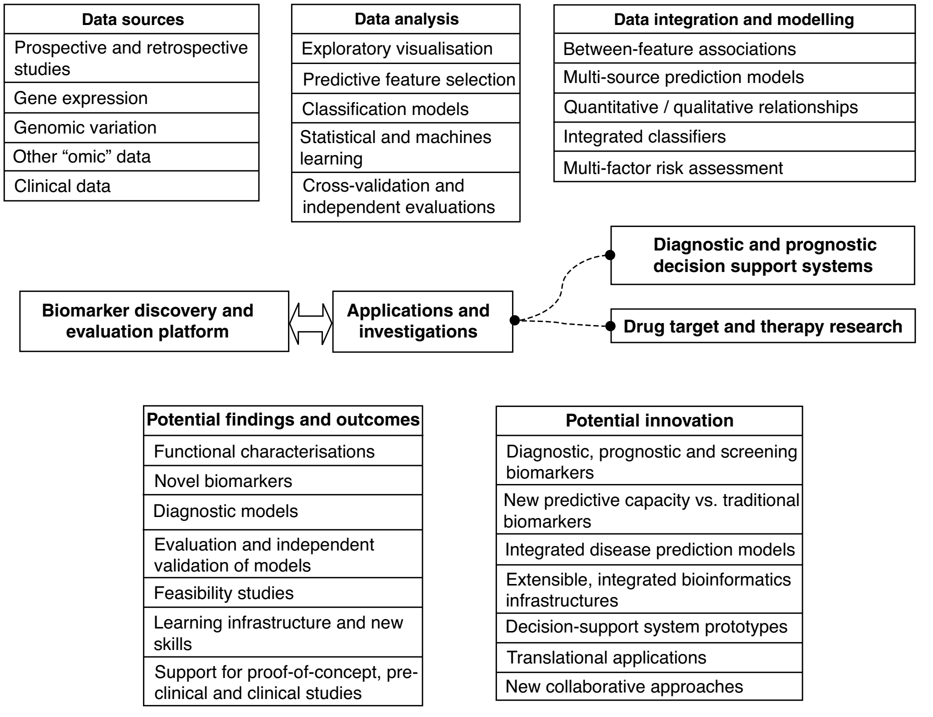


**Figure 9.4** A framework of minimum resources, tools and outcomes required in a bioinformatics infrastructure for supporting biomarker discovery and evaluation

using computational intelligence and advanced information visualization. The potential outcomes facilitated by these infrastructures not only include the acceleration of routine experimental research and the identification of new biomarkers. They are also valuable tools to support training and learning in collaborative, cross-disciplinary and inter-institutional environments. Bioinformatic resources can also be tailored to more specific needs to facilitate translational biomedical research: Proof-of-concept, pre-clinical and clinical studies of new devices, diagnostic kits, decision support systems and therapies. Thus, bioinformatic infrastructures should be seen as backbone engines for the future of translational biomedical research, which demands new and creative ways of integrating resources, people and ideas.

# 10 Challenges and research directions in bioinformatics and biomarker discovery

This chapter will discuss major research challenges and directions for bioinformatics in disease biomarker discovery, with an emphasis on requirements and approaches for prediction model development based on 'omic' data integration and analysis. Major challenges regarding data and information sharing, computational evaluation of biomarkers and prediction models, research reporting practices, research reproducibility and validation of biomarkers and models will be introduced. This chapter also discusses strategies for training researchers in 'translational bioinformatics' and for supporting multi-disciplinary collaboration. Two guest commentaries accompany this chapter to summarize alternative views on problems and challenges, as well as to expand discussions on key computational methods and their applications.

## 10.1 Introduction

It is evident that bioinformatics, generally defined as the development and application of computational technologies for supporting the understanding of biological systems, plays a crucial role in biomedical translational research (Chapter 1). The area of translational bioinformatics focuses on the objective of bridging the gap between

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

biological research and clinical application. This is being accomplished through the development of algorithms, methods and other information resources that bring the bench closer to the bedside. Translational bioinformatics research plays an essential part in disease biomarker discovery, across different clinical domains and using different types of 'omic' data. Personalized approaches to prevent, detect and treat disease using different types of biomarkers and computational tools is a major aspiration in the evolution of modern medicine. These changes also offer promising avenues for identifying more cost-effective and less invasive therapies.

Bioinformatics research supports and drives biomarker discovery investigations that may result in potential novel clinical applications (Bellazzi and Zupan, 2008;Azuaje, Devaux and Wagner, 2009b). For instance, biomarkers can be used to classify patients on the basis of risk categories, which can be used to guide preventive strategies, further screenings or follow-ups. Screening biomarkers and models are used for the early detection of disease in populations of healthy, at high risk, individuals. Also different types of diagnostic applications are being developed to detect disease based on the combination of different types of 'omic' and clinical data. The prediction of responses to treatments or disease progression is another promising application.

Previous chapters reviewed concepts and resources relevant to the design, implementation and evaluation of such applications. Chapter 1 introduced fundamental definitions, problems, requirements and applications of biomarkers. Fundamental statistical and data mining concepts, methods and tools were overviewed in Chapters 2 and 3, which included discussions about technical advantages and limitations of computational solutions and approaches. Chapters 4–8 focused on the analysis and application of different types of 'omic' information in the design and implementation of disease biomarker investigations. Important challenges, technological gaps and requirements for the development of advanced bioinformatics approaches and the exploration of novel clinical applications were presented. Chapter 9 expanded the review of software tools and information resources, which are laying the groundwork for a new generation of computational advances and systems-based approaches.

The greatest obstacles or challenges for bioinformatics and disease biomarker discovery research can be grouped into two major areas: Those challenges that mainly depend on technological changes and innovation, and those that comprise or require major cultural changes. These two areas are clearly intertwined and inter-dependent. On the one hand, changes in scientific culture drive new technological advances. On the other hand, novel computational tools and resources have the potential to redefine research goals and motivations, which in turn may fuel changes in attitudes and culture.

For instance, we will not be able to fully exploit advances in computational biology and bioinformatics without dealing with obstacles relating to open access to information, standards and integrated bio-computing infrastructures. But such obstacles can be eventually overcome by the promotion and development of a culture of sharing in a global, interdisciplinary research environment. Other examples of technological requirements are the availability of more and better data and analysis outcomes to allow research repeatability and greater scrutiny by the scientific community. As discussed below, there is still evidence of the need to improve the rigour of statistical analyses and research reporting practices. The true test of any advance in 'omic' biomarker research will be to pass the filter of rigorous independent evaluations. This also involves comparisons with 'standard' biomarkers or models routinely used in clinical practice on the basis of their classification or prediction performance. Progress in these areas will be constrained, in part, by our capacity to train, mentor and develop a new generation of researchers in translational bioinformatics.

The next sections offer a more detailed discussion of some of the most pressing needs and challenges in bioinformatics for translational biomedical research, in general, and for biomarker discovery research, in particular. The list of problems and challenges covered here complement the data- and method-specific discussions offered in preceding chapters.

## 10.2 Better software

In translational bioinformatics, the success of a software development project may in part be measured by the acceptance of its products and outcomes in a community of users or by their impact in disease-driven research applications. However, above all bioinformaticians should aim to deliver usable software tools that can generate consistent and reproducible results, and that can be easily adapted and maintained in different application contexts. Regardless of the specific requirements and complexity of the tools and their applications, Baxter *et al.* (2006) argue that fundamental practices can be recommended for successful software development: Careful requirement analysis and design before implementation, better documentation of code and processes, the application of quality control in all development phases, the application of community-based standards for representing data and information where possible, and the use of project management tools at all levels.

The typical outcomes of requirement analysis and design not only include documents that synthesize the inputs, processing of the inputs and outputs of the programs, but also specific strategies for testing and evaluating the software. Apart from giving careful consideration to the selection of programming languages and development tools according to specific needs and expectations, it is important to discuss maintenance and release plans. This phase should also consider other quality factors for defining the usefulness and acceptance of the software, such as usability requirements for interface development. Software documentation consists of describing the processes and functions of program modules, as well as in-line comments in the source code. Quality control activities can be performed by ensuring that the programs produce the intended outputs and results consistently, that is software testing, and by tracking and identifying software bugs and potential future conflicts. Developers can ensure that the data inputs and outputs of the software comply with standards for data and information representation, for example XML-based formats for ‘omic’ data. Metadata documents describing the characteristics of the data, encoding formats, definitions and functionality assumptions could be useful when standards are not available for specific types of data or information outputs (Baxter *et al.*, 2006). Different software project management habits, processes and strategies are needed to guarantee that development phases, deliverables and new releases are completed as planned (Berkun, 2005).

Software usability refers to a set of properties that reflect the overall user satisfaction with regard to their capacity to understand the software functionality, as well as the perceived difficulty and complexity of the software. Usability assessment aims to answer questions such as: Are users able to easily obtain information or to generate outputs when they need them? Which difficulties are encountered when trying to complete browsing and navigation tasks? Can users easily learn to interact with the system by inspecting its interface? Recent studies focused on bioinformatic software and users, suggest that usability factors may be hampering the capacity of users in different application domains to effectively and efficiently obtain or process information in their day-to-day research activities. For instance, Bolchini *et al.* (2009) found that browsing- and search-oriented tasks in widely-used Web-based biological databases could be improved by addressing important usability issues. These types of investigations indicate that software designers should carefully consider the compliance of the software functional features with well-defined usability factors relating to content quality and clarity, computing performance, reliability and interface design. Bolchini *et al.* (2009) suggested that some of the biggest challenges in bioinformatics software usability are: (a) users tend to have difficulties when dealing with a long list of results (e.g. gene or document searches), and the visual organization of result items may be directly contributing to this problem; (b) ranking criteria used to organize program outputs may be difficult to understand; and (c) the clarity and quality of the information used to summarize the content of large lists of documents deserve significant improvement.

Other authors have argued that improvement in usability and the adoption of bioinformatics software for large-scale, cross-disciplinary research will be possible through architectures that can support rapid, bottom-up development and integration (Boyle *et al.*, 2008). The latter may comprise the definition of more flexible architectures at a higher level of abstraction to facilitate re-use and adaptation.

## 10.3 The clinical relevance of new biomarkers

Despite the potential of disease biomarkers to improve health, quality of life and survival and the increasing number of publications across different medical domains, there have not been many examples of the successful adoption of novel 'omic' biomarkers in routine clinical practice. The main factor restricting their use and wide acceptance is the lack of conclusive evidence about their clinical relevance, including multi-marker diagnostic and prognostic models.

The demonstration of the clinical relevance of new biomarkers is a critical step that goes beyond the implementation of independent validations (Chapter 1). A clinically-useful biomarker is one that, for instance, can be effectively applied to define the best treatment for a patient, which may in turn improve the quality of life of the patient. The latter can be measured by a reduction of the patient's exposure to invasive interventions or toxic treatments with harmful collateral effects. Apart from aiding in the selection of optimum therapy, a biomarker or prediction model may demonstrate its clinical relevance by allowing an early detection of risk, disease or complications, which can enable a more personalized planning of additional treatments or monitoring strategies (Hinestrosa *et al.*, 2007; Thompson *et al.*, 2008).

An important majority of new biomarkers, based on different types of 'omic' data and computational prediction models, published to date have not produced conclusive evidence of their clinical relevance (Hinestrosa *et al.*, 2007; Contopoulos-Ioannidis *et al.*, 2008; Ledford, 2008; Wang, 2008). A significant proportion of those discoveries have not even passed the threshold of independent replication and validation. Even in the case of independently validated biomarkers that have already influenced the way treatment selection is performed, for example the oestrogen receptor and HER2 (human epidermal growth factor receptor 2) biomarkers for breast cancer, important factors such as standards for measuring the biomarkers and reporting results can represent major obstacles to their widespread acceptance and application (Hinestrosa *et al.*, 2007). Chapter 5 offers examples of multi-gene biomarkers approved for clinical use.

Bioinformatics has an essential role and responsibility in the assessment of the clinical relevance of new disease biomarkers. This can be achieved not only by providing bio-computing infrastructures, tools and advanced computational methodologies, but also by ensuring that such contributions comply with quality standards and best practice for evaluating, reporting and documenting research findings (Section 10.7). Also there is a need to foster methodological rigour and transparency and to strengthen the idea of open collaboration and sharing of resources (e.g. data and well-documented computational models). Moreover, employers, mentors and funding organizations should provide researchers with support and recognition for their commitment to high-quality peer-review, as well as for their active involvement in community-based initiatives designed to meet the challenges described here.

## 10.4 Collaboration

In the biological and clinical sciences there is little doubt about the importance of bioinformatics to support the discovery and development of novel disease biomarkers. Perhaps what is less clear, for some stakeholders in the public and private research sectors, is the capacity or potential that computational researchers can offer to lead knowledge discovery. Is this only about the provision of informatics and data analysis services on demand? Is there widespread recognition for computational research as a key force in the search for new knowledge and in the creation of new biomedical-relevant research directions? Nevertheless, (almost) no one disputes the evidence that inter-disciplinary collaboration is essential for ensuring the successful identification of new technologies for detecting and treating disease in the post-human genome era, independently of the specific roles and opportunities assigned to their participants.

Mutual benefit and trust are essential guiding principles in any collaboration. Computational scientists can contribute to the generation of answers to fundamental biomedical questions through the insights, perspectives, predictions and interpretations obtained from the application of advanced algorithms and methodologies. Moreover, this can drive the formulation of new questions for the consideration of life and clinical scientists. These can be questions connected to the definition of strategies for generating or validating new hypotheses, as well as for prioritizing biomarkers and prediction models according to their potential clinical relevance. But in the end these and other cultural changes needed to improve collaboration and the generation of new knowledge will also depend on the capacity of all researchers, independently of their ‘original background’, to embrace new ways to define and interpret problems. And this may demand from all of us to abandon ‘old disciplines’ or mindsets for the creation of new ones (Eddy, 2005).

## 10.5 Evaluating and validating biomarker models

The main objectives of the evaluation of a disease biomarker model are: (a) to demonstrate its computational and (potential) biomedical relevance and, (b) to estimate "the true" diagnostic or prognostic ability of the model. The standard approaches to achieving these objectives are model cross-validation and independent validations (Azuaje, Devaux and Wagner, 2009b). Different quantitative quality indicators can be applied, such as accuracy, sensitivity and AUC values (Chapter 2).

Model cross-validation involves the generation of disjoint, randomly selected training and testing datasets, which are required for model building (training) and testing independently (Chapter 3). This data sampling process is repeated several times using different training-test partitions, and a global indicator of model quality performance, for example classification accuracy, can be computed by combining the performance indicators derived from the different testing datasets. This evaluation principle, as well as its different versions, has become an accepted practice to offer a more objective, exact and less biased estimation of prediction or classification capability. However, it is also important to remember that cross-validation procedures may overestimate performance in the presence of small datasets (<50 samples) (Braga-Neto and Dougherty, 2004; Molinaro, Simon and Pfeiffer, 2005), or when it is not properly implemented in specific contexts, for example selection bias (Chapter 3) in wrapper-based feature selection and classification.

An independent validation refers to the evaluation of the disease biomarker model using new testing datasets (Chapter 2). The data should be obtained from biological samples and patients that were not previously used during model building or cross-validation (Azuaje, Devaux and Wagner, 2009b). An accepted approach is to have different, independent research teams involved in the sample acquisition, data generation and analysis tasks. Recent research has highlighted the importance of implementing multi-centre studies as a basic condition for independent validation (Ghosh and Poisson, 2009).

Recent investigations have motivated misinterpretations in terms of the potential improvement that newly discovered biomarkers may add to the detection of diseases or clinical responses (Smulders, Thijs and Twisk, 2008). For instance, the INTERHEART study (Yusuf *et al.*, 2004) is a large control-case project (more than 15 000 cases from 52 countries) that concluded that nine conventional physiological risk factors can be used to explain 90% of the cases of cardiovascular disease. A possible misinterpretation is to assume that this result would leave little room for improving cardiovascular disease risk assessment through the incorporation of emerging 'omic' biomarkers (Smulders, Thijs and Twisk, 2008). First, it is important to clarify that new risk determinants can contribute to a better understanding of the development of the disease independently of the predictive power that it may add to an existing model. Risk determinants could be either causal factors or indicators of early disease development. Second, there are many other combinations of risk factors (including conventional or novel, or both) that could achieve similar predictive performance as that reported by the INTERHEART study. Third, it is important not to overemphasize the optimization of single indicators of classification quality (e.g. the ROC AUC, Chapter 2). Although the AUC is a useful and robust indicator for estimating classification performance, several authors have criticized its assumptions, constraints and misuses. For example, Pepe, Cai and Longton (2007), Smulders, Thijs and Twisk (2008) and others have suggested that indicators such as sensitivity, specificity, model calibration measures, and precision should also be considered according to the specific clinical context under study, for example classification or risk prediction applications. Other authors have reported the theoretical constraints (e.g. maximum values) of the AUC in domain- and population-specific analyses. For example, it has been reported that in some applications and populations the theoretical maximum AUC value of an optimal prediction model could be 0.80 (Smulders, Thijs and Twisk, 2008).

In contrast to these difficulties, recent research has also demonstrated how risk assessment and disease detection models based on multiple biomarkers can outperform established models based on traditional risk factors. One example was provided by Zethelius *et al.* (2008), who reported the evaluation of multiple biomarkers for death risk assessment caused by cardiovascular disease in elderly patients. In comparison to the predictive performance offered by individual risk factors commonly applied in the clinical setting, they showed how the combination of multiple biomarkers can improve death prediction performance based on Cox-based models (Chapter 3) and different evaluation indicators. The success of this study can be explained by different design factors (de Lemosand Lloyd-Jones, 2008), such as the inclusion of powerful individual biomarkers with validated roles in cardiac and renal damage, and the careful selection and matching of patients (e.g. inclusion of white males of similar age only). Moreover, in comparison to the majority of previous research in this and other biomedical areas, this study applied a more comprehensive set of prediction performance evaluation indicators, which cover different aspects: model fit, discrimination, calibration and reclassification properties of the prediction models (de Lemos *et al.*, 2008; Zethelius *et al.*, 2008).

On the other hand, it is necessary to recognize that many of the claims about new diagnostic and prognostic biomarkers reported in the literature may be exaggerated and unjustified, and may lack adequate computational evaluations and independent experimental validation (Ioannidis, 2007a; Contopoulos-Ioannidis *et al.*, 2008).

Therefore, there is a need to continue improving our capacity to assess the scientific quality and clinical relevance of new disease biomarker models. An accurate picture of the scientific quality landscape in disease biomarker research can be very difficult to obtain because of the lack of shared evaluation practices, the great diversity of data sources and a publication bias culture. For example, there is empirical evidence suggesting the widespread existence of study (or outcome) publication bias in the scientific literature. That is, research that reports positive, ‘statistically significant’ results or large size effects are more likely to be published (Dwan *et al.*, 2008).

Thus, it has been suggested that the grading of disease biomarker evidence should consider a more integrated, context-dependent view and interpretation of different ‘credibility factors’. For example, Ioannidis (2006) recommended the careful examination of the following factors: Effect size, amount and reproducibility of evidence, protection from bias, biological credibility and clinical relevance. Very small effect sizes (e.g. relative risks or odd ratios $<1.2$) are very unlikely to represent relevant associations in different research environments or clinical settings, even if such relationships or differences are found to be ‘statistically significant’. The credibility of a discovery could be augmented if the obtained effect sizes or associations can be replicated in several independent studies, even in the case of small to moderate sizes. In other situations, large effects may simply reflect the presence of strong bias (e.g. demographic or other phenotypic stratifications) in the population sample under investigation. The detection of possible sources of bias in published studies may be complicated because many research papers lack enough information about the characteristics of the population sample, including clinical information that may explain the effect sizes observed. This is one of the reasons to support efforts to improve scientific reporting practice (Section 10.7). The biological credibility of a study can be established based on available qualitative knowledge and empirical evidence, that is based on how a new finding fits or changes a previously accepted understanding or logic. This factor is typically addressed in the discussion and conclusion sections of research articles. The last critical factor stressed by Ioannidis (2006) refers to the estimation of the clinical relevance of the findings, which was introduced above.

Despite the diversity of requirements, goals and resources available for the development of biomarker discovery and applications, it is possible to define other important common factors to further characterize the usefulness of novel biomarkers in translational medicine. One of them is the development of good practice in evidence reporting (Section 10.7). This does not only refer to the application of sound, consistent experimental and analytical protocols, but also to the detailed and transparent documentation of the different development phases. This also requires a clear and comprehensive specification of resulting biomarkers and computational prediction models. This should not induce inconsistent interpretations and should provide sufficient information to allow the reproducibility of experiments and computational models (Ioannidis *et al.*, 2009). Selective or preferential reporting of 'statistically significant' results should also be avoided (Ioannidis, 2007b).

The clinical efficacy and effectiveness of a new biomarker refers to its potential to benefit patients in different clinical environments (also see Section 10.3). Disagreements exist on whether or not these requirements should be assessed through clinical trials. In some cases, for example the Oncotype DX cancer biomarker system (Chapter 5), approval for clinical application may be granted on the basis of validations using retrospective data. Randomized clinical trials are currently under development for a handful of biomarkers, which mostly fall into the cancer prognostics area. Moreover, it is also important to acknowledge that an important proportion of diagnostic tests applied in day-to-day clinical practice were not assessed in a traditional clinical trials framework (Ioannidis, 2007b). Although it may not be possible to demonstrate the impact of a prediction model on problems such as patient survival, some of these models may contribute to the reduction of unnecessary treatments and the reduction of collateral risks and drug toxicity. Nevertheless, it has been suggested that clinical trials may be useful to estimate the potential effectiveness of biomarkers in specific clinical environments and populations. Proof-of-concept studies, additional retrospective evaluations and carefully designed prospective studies in relatively small populations are also options recommended for assessing the potential clinical efficacy and effectiveness of new biomarkers. These and other options should be seen as the natural continuation of the different evaluation and validation strategies discussed above.

These factors also deal with the question of whether or not a novel biomarker system could be fully deployed, maintained and sustained in a clinical environment. The capacity of a hospital unit to obtain (and process) samples and implement computational predictions in an integrated and efficient fashion are fundamental (Hinestrosa *et al.*, 2007; Ioannidis, 2007b). A great challenge is to assess the capacity of such environments to reproduce the predictive quality reported in previous development phases. Other major challenges involve the training of new staff and the definition of objective standards for quality assessment in specific environments.

The cost-effectiveness of a biomarker system is based on its cost, the overall savings gained in clinical testing procedures, the prevention of unnecessary harms or complications, its support for refining screening strategies, and the reduction of costs incurred by unnecessary treatments (Ioannidis, 2007b). Also it has been suggested that the cost-effectiveness of new biomarkers may be estimated on the basis of their potential as 'modulators' of therapeutic interventions (Ioannidis, 2007b).

## 10.6 Defining and measuring phenotypes

The accurate and unambiguous definition of phenotypes (e.g. disease, prognostic outcomes, conditions, responses to treatments) is a critical factor in the design of biomedically-meaningful computational prediction and classification models. More accurate ways to specify and measure phenotypes can reduce the amount of spurious or false positive associations. It will allow a more controlled selection of patients or biological samples and improve the power of statistical and machine learning techniques. It can aid in the specification or identification of new disease sub-groups and between-group relationships. Moreover, standardized and accurate definitions of disease phenotypes can facilitate the interpretation and reproducibility of methods and results.

In some biomedical areas the assignment of control-disease categories may still be seen as a relatively subjective and non-standardized activity. This is because within the same biomedical discipline there may be a variety of strategies that can be applied to define clinical conditions or specific outcomes. This may depend on the protocols, instruments or analyses (including those that rely on visual inspection) implemented. These concerns are being addressed through the application of more advanced technological platforms for identifying or estimating phenotypes, such as more sophisticated imaging techniques, *in vivo* cell-specific assays and microscale analytical systems. In parallel, different international initiatives will continue to foster the development of standardized biomedical vocabularies, phenotype (non-molecular) data and knowledge bases, together with automated tools for integrating and mining these resources. Moreover, there is room for new community-driven methodologies and resources to facilitate a more exact definition or identification of phenotypes, which may be useful across different biomedical research fields (Freimer and Sabatti, 2003).

## 10.7 Documenting and reporting biomarker research

Researchers may be tempted to assume that the listing of the genes defining a diagnostic biosignature, the summarized description of standard experimental protocols or the presentation of statistical 'significance' probability values are sufficient to ensure an accurate reporting of disease biomarker research. The reporting of disease biomarker investigations requires a clear, rigorous and detailed specification of different qualitative and quantitative aspects, which should enable the reproducibility of computational models and findings in different application settings or using datasets independently generated.

Good reporting practices are required to help researchers, regulators and investors to focus their efforts and resources on only those biomarkers and associated techniques showing clinical relevance potential. By preventing premature (or unnecessary) efforts and investments, researchers also contribute directly to the advancement of public trust in scientific research. Insufficient or inaccurate reporting of results (including negative findings), technology implementation and evaluation may motivate major misrepresentations or misinterpretations of scientific research. An adequate reporting of biomarker research can also enhance the quality and fairness of the peer-review system. Additionally, it creates disincentives to make exaggerated claims and biased interpretations of quantitative analysis results.

Several independent investigations (Sterneand Smith, 2001; Jafari and Azuaje, 2006; Ioannidis, 2007a, 2007b) have pointed out common methodological errors or misinterpretations in studies presenting quantitative models and outcomes. A common problem is the misuse of the expression 'statistical significance' (Chapter 2). Despite advances in the specification of community-driven guidelines for reporting biomarker research, there is still a need to improve the quality of the documentation of model implementations and evaluation results. There is evidence to suggest that a significant number of disease biomarker models, for example those based on gene expression signatures, lack sufficient information to facilitate their external validation or re-implementation (Kostka and Spang, 2008). Even a detailed specification of computational model learning parameters may not provide enough information for defining a biomarker signature unambiguously. This may be explained in part by the incomplete reporting of data pre-processing procedures (e.g. normalization and data transformation algorithms), tools used for model implementation and details about the cross-validation approach applied. Furthermore, it is widely accepted that the selection of normalization techniques may influence classification and prediction (Jafari and Azuaje, 2006; Kostka and Spang, 2008) depending on the data and the application context.

Computational scientists co-authoring research on disease biomarkers should ensure that more careful attention is given to the reporting of software tools, statistical assumptions, hypothesis-testing procedures, potential sources of model over-fitting and the results derived from comparative analyses involving reference or benchmark models. It is crucial to prevent the over-use and misuse of statistical terminology or concepts, such as 'significant' and 'not significant'. Their use should at least be supported by statistic scores and estimated probability values to allow researchers to make their own interpretations. This can also be accompanied by information on assumptions, background knowledge and the potential implications of the differences or relationships found to be statistically detectable. Bioinformatic researchers should also foster an adequate reporting of evidence to show that computational prediction models were trained and tested correctly based on standard data sampling and cross-validation procedures (Chapter 3). This is also related to the problem of assessing potential sources of model selection bias (Chapter 3) (Wood, Visscher and Mengersen, 2007). These challenges and responsibilities should be seen as fundamental to our roles as authors, advisors or peer-reviewers.

The reporting of qualitative or interpretative aspects, as well as those related to the organization of publications, also deserve careful consideration. For instance, many journals have started to encourage authors to report the main limitations of biomedical investigations, as part of their publication guidelines or even as specific formatting instructions in these journals. An improvement in the acknowledgement of the limitations of methodologies and results goes beyond the presentation of the 'future work' to be done. This could also involve a critical discussion of the influence of potential error sources and possible problems in the design and evaluation of models. The need to further improve the quality of reporting of limitations has been stressed in different areas of the life and physical sciences. For example, a recent survey of articles from journals that received the highest impact factors in 2005 indicated that less than 20% of the articles discussed the potential limitations of their studies (Ioannidis, 2007a, 2007b). Also similar concerns have been recently raised in the area of cardiovascular disease biomarkers research (Azuaje, Devaux and Wagner, 2009a). Moreover, it has been shown that many studies in this area can be improved in terms of information completeness and clarity, and in relation to the application of more rigorous quantitative evaluations and interpretations (Azuaje, Devaux and Wagner, 2009a).

Different international, community-driven initiatives have produced reporting guidelines relevant to disease biomarker research and translational bioinformatics. These guidelines aim to address some of the concerns relating to the coverage, depth and quality of information reported. These recommendations represent a valuable source of advice on 'what' and 'how' to report biomarker research in scientific publications. They are applicable to a wide range of biomedical domains and research goals (The Equator Network, 2009): biomarker research for diagnostic and prognostic applications, clinical trials and meta-analyses. In general, such guidelines or 'standards' promote a more accurate, detailed and structured presentation of information (Azuaje, Devaux and Wagner, 2009a).

These and related projects across different application domains have assisted authors and reviewers in improving the readers' confidence in the quality and potential applications of published investigations. The adoption of some of these guidelines and their endorsement by journals have shown to improve the quality of reporting practices and a better understanding of research findings (Smidt *et al.*, 2006a, 2006b).

Examples of guidelines with direct relevance to disease biomarker discovery research are: Consolidated Standards of Reporting Trials (CONSORT), Standards for Reporting of Diagnostic Accuracy (STARD), and Recommendations for Tumour Marker Prognostic Studies (REMARK). CONSORT is a pioneering community-driven initiative tailored to the reporting of clinical trials (The CONSORT Statement, 2009), which has been endorsed by many journals and international publishing organizations.

The STARD focuses on the reporting of disease biomarkers (Bossuyt *et al.*, 2003), and has been endorsed by more than 200 scientific journals since its publication (STARD Statement, 2009). STARD specifies a checklist for reporting diagnostic studies and recommendations for describing study design and development using graphical flow diagrams. The checklist gives specific recommendations on the technical content of the typical main sections of a research paper, for example methods, statistical methods, and results. The REMARK guidelines are an example of an international initiative for improving the reporting of prognostic biomarker research in scientific journals (McShane *et al.*, 2005). These guidelines were recommended for aiding in the evaluation of the clinical relevance of tumour biomarkers with prognostic applications. REMARK offers specific recommendations on the type of content and level of detail required in the different typical sections of a biomedical research publication.

Other examples of guidelines relevant to disease biomarker research are: the QUADAS tool (Quality Assessment of Diagnostic Accuracy Studies) (Whiting *et al.*, 2003), the QUORUM (Quality of Reporting of Meta-analyses) guidelines (Moher *et al.*, 1999) and MOOSE (Meta-analysis Of Observational Studies in Epidemiology) (Stroup *et al.*, 2000). Additional projects and guidelines are described on the EQUATOR Network (2009) Web site. The EQUATOR network aims to improve the quality and reliability of the health research literature by promoting good reporting practices and through the dissemination of guidelines, resources and training activities.

The reporting and documentation of disease biomarkers research will be further augmented by expanding the involvement of computing science and statistical researchers in peer-review and editorial activities. Empirical evidence to support this idea has been provided by a recent randomized trial, which showed that the inclusion of reviewers with a relatively solid background in statistics can improve the quality of research manuscripts presenting diagnostic and prognostic applications (Cobo *et al.*, 2007).

## 10.8 Intelligent data analysis and computational models

The most widely-applied computational methodologies for disease classification or outcome prediction have been based on Cox (proportional-hazards) models and different versions of logistic regression. The former technique is probably the best-known approach to building systems for prognostic studies (survival analysis, Chapter 2) and the prediction of therapeutic responses. The latter technique has been the main approach to diagnostic system development using different types of 'omic' data. However, as Chapter 4–8 and some of the guest commentaries have shown, more advanced methodologies based on different techniques originating from statistical learning and computational intelligence are being explored for different classification problems and biomedical domains (Fogel, Corne and Pan, 2008). For instance, applications based on support vector machines, random forests, neural networks and Bayesian models are becoming more visible in the disease biomarker research literature (Azuaje, Devaux and Wagner, 2009b).

Statistical and machine learning algorithms, such as linear discriminant analysis, different versions of instance-based learning classifiers, variations of decision tree models, neural networks, naïve Bayesian classifiers and support vector machines have been investigated in comparative assessments mainly in the area of gene expression-based biomarker classification (Chapters 3–5). Apart from the studies discussed in Chapter 3, other authors have aimed to compare the predictive or classification capabilities of these techniques in an application-independent context or across different classification scenarios (Baek, Tsai and Chen, 2009). Most of these comparisons have been mainly carried out using different types of public datasets, with an emphasis on samples obtained from cancer studies. Different authors, such as Dudoit, Fridlyand and Speed (2002) and Wang *et al.* (2006 ), have shown that relatively simple (in terms of mathematical complexity) techniques, for example methods based on nearest neighbours, can perform as well or better than more sophisticated algorithms, for example support vector machine classifiers. On the other hand, other studies (Lee *et al.*, 2005;Baek, Tsai and Chen, 2009) have concluded that relatively more complex classifiers, together with different versions of wrapper-based feature selection approaches (Chapter 3), can provide the highest classification performance across different data types. It is also unlikely that a consensus on the predictive capacity of different feature selection techniques will be reached (Saeys, Inza and Larrañaga, 2007).

Such a diversity of observations and suggestive evidence should be expected. The main reason can be simplified in a single word: Context. Specific requirements, constraints and goals of the problem investigated should be the main guiding criteria for model selection, implementation and evaluation. Additionally, until now the discrimination capacity of classification models, as measured by standard performance metrics, have been probably overemphasized as the key criterion to estimate the biomedical potential and computational relevance of new biomarkers and algorithms. In close cooperation with bio-scientists and clinicians, computational intelligence and data mining researchers developing applications for biomarker-based classification and prediction models should consider other quality factors in a more context-dependent and integrated fashion. Examples of other important quality dimensions in a translational research context are: robustness and stability of the biomarker sets selected, biological meaning and relevance of the biomarkers and models to explain underlying mechanisms and prediction outcomes, reproducibility of models and prediction performance in different application settings (Ioannidis *et al.*, 2009), and the potential clinical relevance of biomarkers and models (Sections 10.3 and 10.5). Finally, and independently of the application domain and specific requirements, any solution or approach to meeting these challenges will be constrained by common factors, such as insufficient amounts of data, the curse of dimensionality, lack of adequate documentation of clinical variables, unknown confounding factors, incomplete and evolving knowledge, and major (cross-disciplinary) cultural differences in relation to what 'relevance' may mean in a specific translational research setting.

## 10.9 Integrated systems and infrastructures for biomedical computing

Important goals of translational bioinformatics and disease biomarker discovery will not be accomplished in the absence of advanced, integrative computational infrastructures. These infrastructures should allow researchers to go beyond the 'single-biomarker' research paradigm, and should consist of tools capable of performing data access and analysis in a more integrated and user-friendly fashion (Azuaje, Devaux and Wagner, 2009b). The latter may also require the automated implementation of workflows and intelligent explanations of findings and predictions.

Chapter 9 introduced examples of current initiatives for the development of biocomputing infrastructures, platforms and tools for translational bioinformatics and disease biomarker discovery. A recent example of a national, longer-term initiative is the US National Centres for Biomedical Computing (NCBC, 2009), under the NIH Roadmap for Bioinformatics and Computational Biology. These and other efforts being developed elsewhere (Marcus, 2008) are opening new possibilities for the development of integrated and open software tools more suitable to disease biomarker discovery and translational bioinformatics. In comparison with previous application-independent solutions, the next generation of 'generic' bioinformatics tools will explicitly consider the evolving nature of user and domain requirements. This means that advanced tools will be more flexible, extensible and both user- and developer-friendly, which may facilitate their adaptation to different requirements and constraints. In the long term, this will further facilitate the integration and evolution of new tools and services to accelerate translational research. Furthermore, the open access to integrated bio-computing infrastructures, including high-performance computing resources, will make room for new integrative approaches to discovering and validating disease biomarkers and 'druggable' molecular targets.

## 10.10 Open access to research information and outcomes

The open access to data, metadata, software and computational models and their documentation in sufficient detail is important to support the goals of disease biomarker research as a key engine for the advancement of personalized medicine. This also comprises a sufficient and accurate annotation of biological specimens and samples, and any evidence that can be used to assess the computational and clinical relevance of the biomarkers under investigation.

The availability of well-annotated samples, including complete and accurate description of clinical variables and phenotypes, can also help researchers to group patients in more meaningful phenotype categories and to identify potential sources of population stratification. These tasks are fundamental to increase statistical power and reduce the risk of detecting spurious or confounded associations (Chapter 2). Apart from their support for open source software and open access to publications, the computational biology and bioinformatics communities could expand and coordinate efforts to harmonize practices or criteria for making data, models and other computational resources more accessible.

Also it is expected that more research funding agencies and journals will require researchers to deposit their 'omic' datasets and models into international repositories. In the case of public funding organizations, this is also involving support for open access publishing and requests to make papers freely available after a specific period of time.

Many of the software tools for computational biology and biomarker discovery are freely available or are the product of ongoing open-source software projects (Chapter 9). Relevant examples are the SAM technique (Tusher, Tibshirani and Chu, 2001), the R and Bioconductor projects (Gentleman *et al.*, 2005), Weka (Witten and Frank, 2005) and Cytoscape (Shannon *et al.*, 2003). Despite the prevalence of this culture of sharing in the bioinformatics community, research in disease biomarkers will also depend on a more open exchange of other types of resources and tools originating from other areas, such as clinical informatics. This may comprise vocabularies, text-mining applications and (de-identified) clinical data (Butte, 2008b). The latter should include sufficient associated metadata or descriptions of methodologies to allow the replication of the original analyses and outcomes by external, independent researchers.

## 10.11 Systems-based approaches

The global, integrative characterization of biological systems at different levels of organization based on the combination of different experimental 'omic' and computational approaches, that is systems biology, is already providing new insights into the relation between molecular mechanisms, environment and disease. Over the past five years, important progress has been made in the development of new techniques, tools and theoretical findings that can guide the search for novel disease biomarker and drug targets. Chapters 7 and 8 presented examples of the application of different network-based and integrative data analysis approaches to discovering potential clinically-relevant biomarkers and associations between genes, health, diseases or treatment responses.

Notwithstanding these advances, the success of the journey of systems biology from bench and workstation to 'bedside', that is a systems medicine (Auffray, Chen and Hood, 2009), may take significant time to be realized. Some authors have suggested that, at least for clinicians and the pharmaceutical industry, the deluge of 'omic' approaches piled-up to represent a new era of systems biology is simply the 'culmination of all delusions' (Henney and Superti-Furga, 2008). This scepticism has perhaps been justified by a relative lack of concrete results to suggest the usefulness of systems approaches to support the development of new biomarkers, targets and drugs, or by the impracticality of many of the proposed approaches.

Other authors have suggested that the goals of systems medicine will have a better chance of being achieved by enhancing the coupling of integrative data mining techniques with the dynamic modelling and simulation of complex disease processes beyond their representation as statistic, tissue-independent systems (Vodovotz *et al.*, 2008). This is motivated in part by the need to recognize diseases as dynamic processes, and not as specific states. A system-based approach to understanding disease as an evolving process will require physicians to move beyond the 'single-point in time' paradigm that has been traditionally applied to identify and treat disease (Liebman, 2005). Also, this combination of computational and mathematical knowledge should be driven by the integration of inputs and expertise from diverse clinical and biological fields all the way along the research development cycle, including the design and evaluation of software tools (Section 10.2). One of the key challenges will be to improve the quality and reliability of automated data and information processing tasks that represent the core or starting points of many systems biology investigations. One such task is the construction and refinement of network models of protein-protein (or gene-protein) interactions and signalling pathways, which still demands a significant amount of semi-automated annotation and extraction of information from the scientific literature.

The key to resolving some of these challenges may be a more flexible and context-dependent combination of 'bottom-up' and 'top-down' approaches (Liebman, 2005; Noble, 2006b), that is a more balanced integration of engineering and 'traditional science' approaches. The bottom-up approach to understanding problems and systems is based on the acquisition of data and the search for potentially novel patterns and knowledge in the data. On the other hand, a top-down perspective is accomplished by interrogating the system for clues on its collective function and on the specific role of its most relevant components. The top-down approach is closer to the idea of problem-solving in engineering and requires one to consider a patient as a complex dynamic system (Liebman, 2005).

## 10.12 Training a new generation of researchers for translational bioinformatics

Translational bioinformatics and disease biomarker development will be greatly benefitted by a new generation of computational researchers that can go beyond the design and implementation of databases, algorithms and software tools. They should be given more opportunities to contribute to the formulation of scientific questions together with potential new answers, including new knowledge (i.e. the fundamental goal of computational biology). The diversity of research backgrounds and experiences found in these research areas will further demand a variety of scientific and communicational skills. This also includes a solid understanding of fundamental concepts relating to different biomedical domains, such as: molecular biology, clinical sciences, innovation and research management, bio-ethical issues, research regulation policy, and commercialization of research. However, the development of new research capacities and leadership for translational bioinformatics cannot solely be driven by individual motivation, initiatives restricted to a selected set of research laboratories or short-term efforts. Innovative approaches to 'formal' training, including joint training programmes across universities and departments, may be required to promote a more effective cooperation and to enhance the potential roles of bioinformatics research in truly multidisciplinary research environments (Butte, 2008b). Where joint, cross-departmental training programmes are not available, it will be necessary to expand the level of exposure of medical and biology students to fundamental computational concepts, tools and applications. Similarly, computer science and bioinformatic students should be provided with richer opportunities to investigate disease-driven resources and applications, and with research experiences in the private and public sectors.

Taught and research-driven graduate courses for students with a primary background in biology or medicine will need to further support the development of problem-solving, mathematical and software development skills. This will contribute to the formation of better users of bioinformatics tools and technologies, as well as to the development of potential research partners in the design and evaluation of computational systems and techniques. Different instances of the benefits of research training programs for undergraduate students are also starting to be better known (Taraban and Blanton, 2008). Such programs can have a significant positive impact on the potential capacity of future researchers to identify problems, design solutions and critically communicate findings.

Apart from expanding training and multi-disciplinary contact opportunities, it is important to encourage bioscientists and clinicians to embrace more open-minded attitudes towards bioinformatics-driven research (Azuaje, Devaux and Wagner, 2009b). Similarly, researchers with stronger computational expertise should be allowed to have a more active part in decision-making across all the phases of translational research, including study design and hypothesis-generation tasks. Computational scientists will continue making researchers from other disciplines more aware of the relevance and opportunities offered by rigorous research of new algorithms, methods and tools. A major challenge is to change the perception that many life and medical scientists have about computational biology and bioinformatics as a 'number crunching' activity or as a mere provider of data analysis services on demand.

## 10.13 Maximizing the uses of public resources

The public availability of large collections of 'omic' datasets provides computational biologists with more opportunities to formulate new questions relating to fundamental biomedical problems, as well as to generate potentially novel insights into the occurrence and progression of disease. Amongst such data repositories, GEO (Barrett *et al.*, 2007), ArrayExpress (Brazma *et al.*, 2003), the NCBI Database of Genotypes and Phenotypes (dbGaP) (Mailman *et al.*, 2007), and PRIDE (Martens *et al.*, 2005) have become important resources for depositing and sharing 'omic' data from published research. GEO and ArrayExpress offer gene expression from hundreds of thousands of biological samples and their sizes and content quality are expected to continue to increase. PRIDE and dbGaP are Web-based databases that allow researchers to share mass spectra (Chapter 6) and genome-wide association studies (Chapter 4) data respectively. A review of these and other types of bioinformatic resources was given in Chapter 9.

These open-access data resources are also starting to provide the basis for new integrative approaches to disease biomarker discovery and predictive modelling. For example, Camargo and Azuaje (2008) combined different GEO datasets, information from annotated protein-protein interaction networks and machine learning techniques to proposed potentially novel biomarkers of dilated cardiomyopathy in humans. English and Butte (2007) combined publicly-available datasets from gene expression, proteomics and RNAi experiments, and identified known and potentially new associations between some of these genes and obesity. As in Camargo and Azuaje's investigation, they demonstrated that prediction models based on the integration of multiple datasets can outperform models built on single-source datasets independently. Chapters 7 and 8 review network-based and other integrative data analysis approaches to disease biomarker discovery.

## 10.14 Final remarks

Although this chapter, and the book as a whole, simply skims the surface of a large and diverse collection of problems, challenges and opportunities, it is still possible to present general conclusions and recommendations for guiding research in translational bioinformatics and disease biomarkers. Table 10.1 summarizes key challenges and potential research directions for bioinformatics and disease biomarker discovery. Alternative or more detailed discussions and recommendations, including those tailored to specific biomedical areas, can be obtained in (Bellazzi and Zupan, 2008), (Thompson *et al.*, 2008) and (Azuaje, Devaux and Wagner, 2009b).

These challenges and recommendations are a reflection of the complexity and costs associated with the discovery of potentially relevant disease biomarkers. Although an exhaustive discussion of techniques, problems and applications are constrained by time and publication space, we hope that the content of this book can at least offer the reader alternative or broader perspectives for the design, implementation and interpretation of disease biomarker studies. Progress and obstacles should equally provide us with the motivation to envision new possibilities. Possibilities that can enable us to make a difference, for people and the advancement of knowledge everywhere.

**Table 10.1** Key challenges and potential research directions for bioinformatics and disease biomarker discovery

| Challenge | Needs and directions | Recommended additional reading |
|---|---|---|
| Software development | Reliability, rapid adaptability and flexibility; 'bottom-up' development approaches; open-source and extensible solutions; meeting specific biomarker research requirements in industry and academic settings; more attention to usability issues. | (Berkun, 2005); (Baxter *et al.*, 2006); (Boyle *et al.*, 2008); (Bolchini *et al.*, 2009); Chapter 9 of this book |
| Clinical relevance | Clearer definitions and strategies; community-wide dialogue; more research on the role of pilot trials; rigours study design; better reporting practice; extensive cross-validation and independent validation. | (Hinestrosa *et al.*, 2007); (Thompson *et al.*, 2008); Loscalzo (2009) |
| Collaboration | Innovative training schemes for researchers and students; more active participation at the interface between the clinical, life and computational sciences; policy, funding and management actions to promote scientific diversity, new research approaches and communication. | (Eddy, 2005); Vicens and Bourne (2007); Nurse (2008); (Butte, 2008b) |
| Evaluation methods | Correct implementation of cross-validation and independent evaluations; addressing model selection and publication bias; more accurate reporting and documentation of results and models; diverse and application-dependent quality indicators; clinical relevance assessment. | (Ioannidis, 2006, 2007b); (Hinestrosa *et al.*, 2007); (Pepe, Feng and Gu, 2007); (Dwan *et al.*, 2008); (Smulders, Thijs and Twisk, 2008), Chapter 3 of this book |
| Phenotype definitions | More accurate and unambiguous definition of phenotypes; community-based initiatives; biocomputing resources and infrastructure; incorporation of new technologies to measure physiological variables; new semantic and terminological platforms. | (Freimer and Sabatti, 2003); (Butte, 2008a) |
| Documenting and reporting models | Community-based guidelines; emphasis on the reporting of sufficient information to enable reproducibility and independent evaluations; improvement of multi-disciplinary communication and training. | EQUATOR Network (2009) |

**Table 10.1** (*Continued*)

| Challenge | Needs and directions | Recommended additional reading |
|---|---|---|
| Data analysis | Knowledge-driven approaches; computational intelligence; diverse and application-dependent predictive and classification performance assessment methodologies; integrated 'omic' models vs. traditional biomarkers. | (Fogel, Corne and Pan, 2008); (Bellazzi and Zupan, 2008); (Baek, Tsai and Chen, 2009), Chapter 3 of this book |
| Integrated computing infrastructures | Automated implementation of work flows and intelligent explanations of findings and predictions; high-performance, Grid-based platforms; projects demonstrating the value of integrated biocomputing infrastructures in translational research; open access; international funding and cooperation. | (Marcus, 2008); (NCBC, 2009); Chapter 9 of this book |
| Open access to research information and outcomes | Transparent and accurate reporting of findings; tools and standards for data and software sharing; integration of 'omic' and phenotype databases; more participation and funding for open-source and open-access research. | Feller *et al.* (2005); Willinsky (2006); (Butte, 2008b); (NCBC, 2009) |
| Systems-based approaches | Combination of hypothesis- and discovery-drive research, top-down and bottom-up approaches; demonstration of the application of systems biology to develop new clinically relevant biomarkers or identify novel druggable; integrative bioinformatic platforms. | (Loscalzo, Kohane and Barabasi, 2007); (Henney and Superti-Furga, 2008); (Auffray, Chen and Hood, 2009) |
| Training and education | Translational bioinformatics training at undergraduate, graduate and post-doctoral levels; new recognition mechanisms for public service and outreach, mentoring and cross-disciplinary training. | (Butte, 2008b); (Taraban and Blanton, 2008) |
| Maximization of public resource use | Integrative data mining of 'omic' and clinical data repositories; new methodologies for linking genes; processes and phenotypes using multiple, independent sources of data; new research approaches and support for sustaining public resources. | Brazma, Krestyaninova and Sarkans (2006); (Butte, 2008b); (Ioannidis *et al.*, 2009); Nucleic Acids Research database issues (nar.oxfordjournals.org) |

# Guest commentary (1) on chapter 10: Towards building knowledge-based assistants for intelligent data analysis in biomarker discovery

**Riccardo Bellazzi**
*Dipartimento di Informatica e Sistemistica, Università degli Studi di Pavia, 27100, Pavia Italy*

The key challenges and potential research directions for bioinformatics and disease biomarker discovery highlighted in Chapter 10 of this book well describe the most important issues that the field should face in the next few years. Amongst such challenges, a very intriguing one for bioinformaticians is related to data analysis methods to support biomarker discovery. As properly reported by Francisco Azuaje, the extraction of multivariate predictive models based on machine learning and statistical techniques

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

may effectively provide tools for diagnosis and prognosis based on a panel of biomarkers, or ‘omic’ signatures. Since the end of the last century several approaches based on gene expression microarrays (Brown *et al.*, 2000) and on mass spectrometry data have been proposed (Yu *et al.*, 2005). However, several of those approaches suffered from lack of reproducibility; moreover different models with the same prediction capability starting from the same data set were derived. As a matter of fact, this problem is related to both experimental and data analysis pitfalls (Hu, Loo and Wong, 2006). As far as the data analysis problems are concerned, the automated extraction of multivariate models is heavily constrained by the intrinsic limitations of ‘large-m small-n’ data sets that are frequently available in bioinformatics. High dimensional feature spaces are affected by the problem known as the curse of dimensionality. The curse of dimensionality is related to the ‘exponential increase in volume associated with adding extra dimensions to a feature space’ (Wikipedia, 2009); in other words, the number of measurements needed to describe a feature space with the same accuracy increases exponentially with the number of features. In the case of DNA microarrays and mass spectrometry data, the feature space ranges from the order of ten thousands to hundred thousands. This number increases to 1 million in the case of SNP microarrays. Building classification models from such a kind of feature spaces is extremely difficult, as the greatest part of the space will be ‘empty’, that is without any data point, and even the ‘dense’ regions may turn out to be highly under-sampled. Unfortunately, given the sparseness of the data with respect to the feature space, also the feature selection step may be prone to include irrelevant variables. Therefore, a purely data-driven approach may have no chances to derive ‘robust models’, even if state-of-art validation approaches are applied, such as cross-validation and bootstrap. This has two main side effects: the classifiers may have large variance and several solutions to the same problem may exist (Mramor *et al.*, 2007), with no guarantee that the selected features have biological relevance. As stated in Chapter 10, the remedy to this problem is using prior knowledge in the data analysis process; the well known Gene Set Enrichment Analysis method is a first answer to these issues (Subramanian *et al.*, 2005). In bioinformatics, the number of knowledge-sources in electronic formats is becoming huge. Together with biological databases, such as GeneBank or SwissProt, a collection of secondary data resources are available, such as Gene or MIPS, as well as knowledge repositories such as the Gene Ontology or OMIM. Another important knowledge source is represented by the medical literature, available in electronic format in PubMed. In the second part of this decade there have been several papers dealing with information retrieval and knowledge mining of the above mentioned resources. The available knowledge may be used in a variety of ways. In feature selection it helps to extract variables related to the biological problem at hand and/or to exclude redundant variables. When the predictive model is learnt, prior knowledge is used to properly combine heterogeneous information through, for example, Bayesian models or ensemble classifiers. It is important to note that the methods that are designed to automatically incorporate prior knowledge in the learning process are able to process, together with the available data (say genes, mass/charge ratios, SNPs, etc.), also the available ‘annotations’, that is the information on the biomedical role and biological nature of the measured features. To this end, it is possible to apply mathematical and probabilistic algorithms that work in the space of ‘annotations’, in order to select multivariate biomarkers also on the basis of their potential usefulness and meaningfulness.

Thanks to the importance of prior knowledge, it is likely that in the future Artificial Intelligence (AI) will provide increasingly important contributions to biomarker discovery. The research area known as 'Intelligent Data Analysis' (IDA) deals with the application of AI methods and tools to data analysis (Zupan, Keravnou and Lavrac, 1997). IDA seems particularly suited to provide ways to automatically incorporate prior information to constrain the search of prediction and classification models in 'difficult' feature spaces. AI approaches, however, will not only enable the data analysis process to be significantly boosted, but also the support tools for the entire knowledge discovery process to be built. A very interesting and challenging issue is to develop software tools that are able at the same time to handle the huge amounts of heterogeneous data produced, have access to the knowledge repositories and eventually facilitate the biomarker selection workflow. Although several tools and workflow systems have been developed to provide an efficient way of handling, integrating, manipulating and exploring biological information, very few experiences exist in the area of supporting the reasoning process underlying scientific discovery (King *et al.*, 2009). To this end, it seems very interesting to apply formal conceptual models as the basis of the design of knowledge discovery support tools. In the late 1980s and early 1990s automated reasoning was widely studied in the area of expert systems in medicine. Amongst others, an epistemological model for scientific discovery, called Select and Test Model (STModel), was proposed and successfully applied to model medical knowledge-based systems (Ramoni *et al.*, 1992). Very recently, Nuzzo *et al.* (2009) defined an instance of the ST-model to provide a conceptual framework for genome-wide studies and to facilitate the development of automated decision support systems for genomic studies.

The ST-Model structure is made of different inference steps: hypotheses generation (or selection), which is divided into abstraction and abduction, while the testing phase consists of hypotheses ranking, deduction and induction. Following this model, it is possible (i) to model reasoning activities which underlie biomarker discovery; (ii) to design and implement tools to assist the discovery process. In their paper, Nuzzo and colleagues limited their analysis to genome-wide association studies. In this case, *abstraction* consists of the definition of a phenotype of interest and the selection of the individuals to be studied. The *abduction* step is performed by running standard statistical association tests, which generate a set of candidate markers associated with the phenotype, and therefore a set of hypotheses to be tested. Such hypotheses are of the kind 'marker x is associated with the phenotype'. The *deduction* step considers each hypothesis in turn and compares it with the knowledge available in knowledge repositories. In particular, the deduction step is aimed at deriving the necessary consequences that must hold in case the hypotheses were true. For example, this step may entail determining whether a certain marker may be related to pathways or Gene Ontology classes that are in some way related to the disease; this may also imply that other measurable variables (say SNPs, genes, proteins) should be expected to be altered in differential analysis. Additional evidence may allow the implementation of an *eliminative induction* step that reduces the hypothesis space or the repetition of the process in a more focused way by redefining the phenotype or changing the initial markers set. Thanks to its high-level conceptualization, the model can be instantiated also in other studies and could be thought of as a basis for a discovery support tool for different kinds of biomarkers (e.g. transcription factor binding sites identification, knock-out gene experiments).

As brilliantly reported by F. Azuaje in Chapter 10, the discovery of biomarkers poses challenges which are strongly interdisciplinary. The development of bioinformatics tools to support this complex task requires a perfect understanding of the biomedical and clinical problems and a tight communication between all actors involved, ranging from clinicians to software developers. Such a goal may be achieved by improving training in translational bioinformatics, designing and implementing new computational infrastructures and, finally, exploiting all methods which allow the representation, sharing and reuse of knowledge in all phases of scientific discovery. The paradigm of expert systems and decision support tools may unexpectedly find a new, important role in the area of molecular biology as enablers of more efficient strategies to unravel the basic mechanisms of life and to help transfer the new knowledge into useful biomedical results.

## References


Brown, M.P., Grundy, W.N., Lin, D. *et al.* (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. *Proceedings of the National Academy of Sciences of the United States of America*, **97** (1), 262–267.

Hu, S., Loo, J.A. and Wong, D.T. (2006) Human body fluid proteome analysis. *Proteomics*, **6** (23), 6326–6353.

King, R.D., Rowland, J., Oliver, S.G. *et al.* (2009) The automation of science. *Science*, **324** (5923), 85–89.

Mramor, M., Leban, G., Demsar, J. and Zupan, B. (2007) Visualization-based cancer microarray data classification analysis. *Bioinformatics*, **23** (16), 2147–2154.

Nuzzo, A., Riva, A., Stefanelli, M. and Bellazzi, R. (2009) An architecture for automated reasoning systems for genome-wide studies. Proc. AIME 2009, Verona.

Ramoni, M., Stefanelli, M., Magnani, L. and Barosi, G. (1992) An epistemological framework for medical knowledge-based systems. *IEEE Transactions on Systems, Man and Cybernetics*, **22** (6), 1361–1375.

Subramanian, A., Tamayo, P., Mootha, V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (43), 15545–15550.

Wikipedia (2009) http://en.wikipedia.org/wiki/Curse_of_dimensionality, [last accessed June 4th 2009].

Yu, J.S., Ongarello, S., Fiedler, R. *et al.* (2005) Ovarian cancer identification based on dimensionality reduction for high-throughput mass spectrometry data. *Bioinformatics*, **21** (10), 2200–2209.

Zupan, B., Keravnou, E. and Lavrac, N. (1997) *Intelligent Data Analysis in Medicine and Pharmacology*, Kluwer, Norwell (MA).

# Guest commentary (2) on chapter 10: Accompanying commentary on 'challenges and opportunities of bioinformatics in disease biomarker discovery'

**Gary B. Fogel**
*Natural Selection, Inc., San Diego, CA 92121, USA*

## Introduction

At no time in human history has there been a better opportunity to understand the basis of disease and offer remediation to the patient. The last 50 years of growth in biotechnology have provided us with an amazing understanding of the cell, the onset of disease, and

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

patient monitoring, all enabling rapid decisions about proper healthcare. Microarrays are commonplace, measuring the expression of thousands of genes simultaneously. New and innovative technologies allow us to monitor the growth of cells in continuous fashion, and technology will soon arrive that allows us to sequence complete human genomes for <$1000. These and other methods drastically increase our capability to explore and examine the processes that lead to disease, and help us map systems biology. However, these same approaches also generate terabytes or petabytes of data with each experiment, each with large dimensions of concern. Thus, it becomes ever more important that along with these new technologies, we also seek to develop new methods of data interpretation to assist with proper data mining and decision making. Nowhere is this more true than with translational medicine, given that biological systems are inherently nonlinear, continuous, and dynamic.

## Biocyberinfrastructure

In light of this requirement for better data collection and mining, the US National Science Foundation has recently recognized the importance of 'cyberinfrastructure' (Atkins *et al.*, 2003) to the future of scientific discovery. The term cyberinfrastructure refers to 'infrastructure based upon distributed computer, information and communication technology. . .required for a knowledge economy' (Atkins *et al.*, 2003). Thus, governments are already aware of the importance of better storage, communication, and information retrieval systems in light of exponential increases in the volume of data collection. While high-throughput data continues to be generated, integration and curation of the data both remain central issues (Stevens, 2006). Others have noted the importance of bioinformatics in accelerating discovery, in a manner thought to be 10 to 100 times more efficient than using wet lab experiments on their own (Stevens, 2006). While wet lab experiments are still critical for evaluation of computational hypotheses, they are likely to remain expensive. Personnel time only adds to this expense. While it is that computational tools provide much higher-throughput solutions to problems, and can evaluate much larger solution spaces than humans, perhaps one of the large hurdles that remains is the fear that computers will replace the need for humans in the decision loop. A reasonable solution is to provide computational tools as an assistant to clinician experts who will be charged with the responsibility of final healthcare decisions. Clinicians are not always eager to see this eventuality.

## Government regulations on biomarker discovery

Another significant challenge to biomarker discovery is government approval. If translational medicine is really to succeed, then government agencies need to understand the latest advancements in technology and be willing to leverage their contribution with proper oversight. In the United States, the field of *in vitro* diagnostics is regulated by the Food and Drug Administration (FDA). The FDA has recently issued guidance on pharmacogenetic and genetic tests for heritable biomarkers, providing the community with an understanding of how the government intends to regulate disease biomarkers that make it to clinical practice. These guidelines are already affecting the way in which biomarkers are being accepted, with, in some cases, a negative effect on sales (Ray, 2009). Moreover, these same guidelines might also affect the way in which the *in vitro* diagnostic models such as artificial neural networks or other types of machine learning approaches are to be accepted in decision support when using multiple biomarkers in combination (Lisboa and Taktak, 2006). Clearly it is beneficial to have effective rules that protect the end-user from alarming rates of false positive or false negative decisions. However, much more needs to be done to raise awareness in the scientific and government regulatory communities about the true utility of these approaches. And while it may be that 'black box' approaches such as neural networks are not easily approved, it may be that these methods can outperform linear methods for predicting outcome from biomarkers. Should we not be translating the best decision processes possible to treat human diseases even if we may not fully understand how these decisions are being reached? The computational intelligence community has a wide range of possible machine learning approaches (and their hybridization) to offer that can provide significant value for healthcare (Fogel, 2006; Fogel, Corne and Pan, 2008). One significant hurdle in promoting the continued development of these computational tools is to avoid needless overregulation while simultaneously protecting clinician and patient rights to understand how healthcare decisions are being made from these models. The validation of medical neural networks has been a long-standing concern that has yet to have an adequate solution (Rodvold, 2001).

## Computational intelligence approaches for biomarker discovery

The field of computational intelligence (artificial neural networks, fuzzy systems, and evolutionary computation), provides tremendous opportunity for rapid identification of novel biomarkers, in combination, for diagnosis or prognosis. With neural networks, biomarkers are treated as features that can be input to a model, and combined in nonlinear combinations to produce an output decision (or set of decisions). The topology of the neural network can either be predefined for the dataset at hand, or the topology of the neural network can itself be optimized through a search of possible neural network topologies using evolutionary algorithms (Fogel, 2008; Lamers *et al.*, 2008). This approach for simultaneous model optimization with feature selection results in rapid learning without a requirement for expert user intervention. The simulated evolutionary process literally searches the model space using variation and selection to arrive at useful models in light of the data at hand. In some cases the features themselves may not be easily discretized. For example, the features may be with respect to a colour such as orange, and it is difficult to know where 'orange' precisely differs from yellow or red. For such problems, a set of fuzzy rules can be generated without requiring the need for arbitrary threshold placement on the data. All of these approaches are now being applied to for biomarker discovery on a regular basis and promise to help revolutionize the way in which large datasets can be evaluated.

## Open source data, intellectual property, and patient privacy

Open-source solutions that help build cyberinfrastructure capabilities are valued highly in both the academic and industrial communities. However, open source can itself provide some barriers and constraints for progress. For instance, in terms of patient data, curators of open source datasets have to carefully remove any information that could be used to tie health data back to specific patients when the data is made available to the public. Open source datasets can contain errors, especially when multiple users have helped generate the data, and constant curation is a requirement. Furthermore, pharmaceutical companies may wish to develop their own datasets after having carefully studied a particular cancer pathway for years, and in competition with other companies interested in similar diagnostics. For these companies, development of intellectual property in the area of biomarker discovery is a critical component of their future success. Open source data and tools are not often viewed by companies as a viable means of achieving a corporate advantage in a marketplace. Thus, another hurdle for translational medicine is the importance of successful advancement in clinical practice with a desire (but not a requirement) for open source data, that still allows for the prospects of a competitive marketplace while simultaneously advancing clinical practice and preserving company intellectual property and patient confidentiality. The importance of open source solutions should not preclude the independent commercialization of technology.

## Conclusions

Biological systems are inherently nonlinear, continuous, and dynamic processes. Identification of disease biomarkers therefore, requires modelling approaches that can model the system appropriately without assuming feature independence. As our ability to generate biological data increases, interpretation of that data only becomes more critical as we attempt to define a true system-level understanding. Translating these discoveries to clinical practice requires appropriate controls for patient safety and confidentiality. The balance of appropriate regulation, in light of useful tools to enhance healthcare, is key to the successful application of modern machine learning approaches such as computational intelligence in this domain. Without a doubt, our understanding of the mapping of genotype to phenotype in light of the environment will increase rapidly over the next decade. It is indeed quite exciting to realize that we stand at the threshold of an entirely new way of translating tremendous volumes of data into better healthcare decisions. However, many significant challenges with data storage and interpretation, model development and regulation, open source frameworks and intellectual and patient privacy continue to affect our ability to produce this outcome. While better computational methods such as computational intelligence approaches are a start, a new level of thinking has to be applied at many levels simultaneously for the translation to occur.

## References

Atkins, D.E., Droegemeier, K.K., Feldman, S.I. *et al.* (2003) *Revolutionizing Science and Engineering Through Cyberinfrastructure: Report of the National Science Foundation Blue-Ribbon Advisory Panel on Cyberinfrastructure*, National Science Foundation, Washington, DC.

Fogel, G.B. (2006) Deciphering mechanisms of gene regulation through evolutionary computation. *IEEE Computational Intelligence Magazine*, **1** (4–5), 40.

Fogel, G.B. (2008) Computational intelligence approaches for pattern recognition in biological systems. *Briefings in Bioinformatics*, **9**, 9307–316.

Fogel, G.B., Corne, D.W. and Pan, Y. (2008) *Computational Intelligence in Bioinformatics*, Wiley-IEEE Press, New York.

Lamers, S.L., Salemi, M., McGrath, M.S. and Fogel, G.B. (2008) Prediction of R5, X4, and R5X4 HIV-1 coreceptor usage with evolved neural networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, **5**, 291–300.

Lisboa, P.J. and Taktak, A.F.G. (2006) The use of artificial neural networks in decision support in cancer: a systematic review. *Neural Networks*, **19**, 408–415.

Stevens, R. (2006) Trends in cyberinfrastructure for bioinformatics and computational biology. *CTWatch Quarterly*, August.

Ray, T. (2009) Lack of FDA guidance on KRAS testing creates confusion; Pharmas see lower sales. *Pharmacogenomics Reporter*, February 4, 2009.

Rodvold, D.M. (2001) Validation and regulation of medical neural networks. *Molecular Urology*, **5**, 141–145.

# References

Abate-Shen, C. and Shen, M.M. (2009) Diagnostics: the prostate-cancer metabolome. *Nature*, **457**, 799–800.

Aerts, S., Lambrechts, D., Maity, S. *et al.* (2006) Gene prioritization through genomic data fusion. *Nature Biotechnology*, **24**, 537–544.

Albert, R. (2005) Scale-free networks in cell biology. *Journal of Cell Science*, **118**, 4947–4957.

Alexandrov, T., Decker, J., Mertens, B. *et al.* (2009) Biomarker discovery in MALDI-TOF serum protein profiles using discrete wavelet transformation. *Bioinformatics (Oxford, England)*, **25**, 643–649.

Al-Shahrour, F., Carbonell, J., Minguez, P. *et al.* (2008) Babelomics: advanced functional profiling of transcriptomics, proteomics and genomics experiments. *Nucleic Acids Research*, **36**, W341–W346.

Al-Shahrour, F., Minguez, P., Tárraga, J. *et al.* (2006) BABELOMICS: a systems biology perspective in the functional annotation of genome-scale experiments. *Nucleic Acids Research*, **34** (Web Server issue), W472–W476.

Alterovitz, G., Xiang, M., Liu, J. *et al.* (2008) System-wide peripheral biomarker discovery using information theory. *Pacific Symposium on Biocomputing*, **13**: 231–242.

Altshuler, D., Daly, M.J. and Lander, E.S. (2008) Genetic mapping in human disease. *Science*, **322**, 881–888.

Ambroise, C. and McLachlan, G.J. (2002) Selection bias in gene extraction on the basis of microarray gene-expression data. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6562–6566.

Anastassiou, D. (2007) Computational analysis of the synergy among multiple interacting genes. *Molecular Systems Biology*, **3**, 83.

Anderson, L. (2005) Candidate-based proteomics in the search for biomarkers of cardiovascular disease. *The Journal of Physiology*, **563** (Pt 1), 23–60.

Arab, S., Gramolini, A.O., Ping, P. *et al.* (2006) Cardiovascular proteomics: tools to develop novel biomarkers and potential applications. *Journal of the American College of Cardiology*, **48**, 1733–1734.

Archacki, S.R., Angheloiu, G., Tian, X.L. *et al.* (2003) Identification of new genes differentially expressed in coronary artery disease by expression profiling. *Physiological Genomics*, **15**, 65–74.

*Bioinformatics and Biomarker Discovery: "Omic" Data Analysis for Personalized Medicine* Francisco Azuaje

Asgharzadeh, S., Pique-Regi, R., Sposto, R. *et al.* (2006) Prognostic signifi cance of gene expression profi les of metastatic neuroblastomas lacking MYCN gene amplifi cation. *Journal of the National Cancer Institute*, **98**, 1193–1203.

Auffray, C., Chen, Z. and Hood, L. (2009) Systems medicine: the future of medical genomics and healthcare. *Genome Medicine*, **1**, 2.

Azuaje, F. (2003) Clustering-based approaches to discovering and visualising microarray data patterns. *Briefings in Bioinformatics*, **4**, 31–42.

Azuaje, F. (2006) Review of "Data Mining: Practical Machine Learning Tools and Techniques" by Witten and Frank. *BioMedical Engineering OnLine*, **5**, 51.

Azuaje, F., Devaux, Y. and Wagner, D. (2009a) Challenges and standards in reporting diagnostic and prognostic biomarker studies. *Clinical and Translational Science*, **2**, 156–161.

Azuaje, F., Devaux, Y. and Wagner, D. (2009b) Computational biology for cardiovascular biomarker discovery. *Briefings in Bioinformatics*, **10**: 367–377.

Azuaje, F. and Dopazo, J. (eds) (2005) *Data Analysis and Visualization in Genomics and Proteomics*, Wiley, London, UK.

Bader, G.D. and Hogue, C.W. (2003) An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, **4**, 2.

Baek, S., Tsai, C.A. and Chen, J.J. (2009) Development of biomarker classifiers from high-dimensional data. *Briefings in Bioinformatics*, **10**, 537–546.

Bair, E. and Tibshirani, R. (2004) Semi-supervised methods to predict patient survival from gene expression data. *PLoS Biology*, **2**, E108.

Balding, D.J. (2006) A tutorial on statistical methods for population association studies. *Nature Reviews. Genetics*, **7**, 781–791.

Barabási, A.L. (2003) *Linked: How Everything Is Connected to Everything Else and What It Means*, Penguin Books, Cambridge, Massachusetts, USA.

Barabási, A.L. and Oltvai, Z.N. (2004) Network biology: understanding the cell's functional organization. *Nature Reviews. Genetics*, **5**, 101–113.

Barla, A., Jurman, G., Riccadonna, S. *et al.* (2008) Machine learning methods for predictive proteomics. *Briefings in Bioinformatics*, **9**, 119–128.

Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics (Oxford, England)*, **21**, 263–265.

Barrett, T., Troup, D.B., Wilhite, S.E. *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles–database and tools update. *Nucleic Acids Research*, **35** (Database issue), D760–D765.

Barth, A.S., Kuner, R., Buness, A. *et al.* (2006) Identification of a common gene expression signature in dilated cardiomyopathy across independent microarray studies. *Journal of the American College of Cardiology*, **48**, 1610–1617.

Basso, K., Margolin, A., Stolovitzky, G. *et al.* (2005) Reverse engineering of regulatory networks in human B cells. *Nature Genetics*, **37**, 382–390.

Baxter, S.M., Day, S.W., Fetrow, J.S. and Reisinger, S.J. (2006) Scientific software development is not an oxymoron. *PLoS Computational Biology*, **2** (9), e87.

BCC Research (2008) [http://www.bccresearch.com].

Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A. (2004) The genetic association database. *Nature Genetics*, **36**, 431–432.

Bellazzi, R. and Zupan, B. (2008) Predictive data mining in clinical medicine: current issues and guidelines. *International Journal of Medical Informatics*, **77**, 81–97.

Bellew, M., Coram, M., Fitzgibbon, M. *et al.* (2006) A suite of algorithms for the comprehensive analysis of complex protein mixtures using high-resolution LC-MS. *Bioinformatics (Oxford, England)*, **22**, 1902–1909.

Benjamin, I.J. and Schneider, M.D. (2005) Learning from failure: congestive heart failure in the postgenomic age. *The Journal of Clinical Investigation*, **115**, 495–499.

Benjamini, Y. and Hochberg, Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, **57**, 289–300.

Bennett, B.J., Romanoski, C.E. and Lusis, A.J. (2007) Network-centered view of coronary artery disease. *Expert Review of Cardiovascular Therapy*, **5**, 1095–1103.

Bennett, L., Palucka, A.K., Arce, E. *et al.* (2003) Interferon and granulopoiesis signatures in systemic lupus erythematosus blood. *The Journal of Experimental Medicine*, **197**, 711–723.

Berkun, S. (2005) *The Art of Project Management*, O'Reilly Media, Sebastopol, CA, USA.

Bernard, S.A., Gray, T.W., Buist, M.D. *et al.* (2000) Treatment of comatose survivors of out-of-hospital cardiac arrest with induced hypothermia. *The New England Journal of Medicine*, **346**, 557–563.

Bild, A.H., Yao, G., Chang, J.T. *et al.* (2006) Oncogenic pathway signatures in human cancers as a guide to targeted therapies. *Nature*, **439**, 353–357.

Biomarkers Definitions Working Group (2001) Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clinical Pharmacology and Therapeutics*, **69**, 89–95.

BIRN (2008) The Biomedical Informatics Research Network, [www.nbirn.net], accessed 18 December 2008.

Blankenberg, S., McQueen, M.J., Smieja, M. *et al.* (2006) Comparative impact of multiple biomarkers and N-terminal pro-brain natriuretic peptide in the context of conventional risk factors for the prediction of recurrent cardiovascular events in the heart outcomes prevention evaluation (HOPE) Study. *Circulation*, **114**, 201–208.

Bodenreider, O. (2004) The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, **32** (Database issue), D267–D270.

Bolchini, D., Finkelstein, A., Perrone, V. and Nagl, S. (2009) Better bioinformatics through usability analysis. *Bioinformatics (Oxford, England)*, **25**, 406–412.

Boser, B.E., Guyon, I.M. and Vapnik, V.N. (1992) A training algorithm for optimal margin classifiers. Proc. of 5th Annual ACM Workshop on COLT (ed. D. Haussler), pp. 144–152.

Bossi, A. and Lehner, B. (2009) Tissue specificity and the human protein interaction network. *Molecular Systems Biology*, **5**, 260.

Bossuyt, P.M., Reitsma, J.B., Bruns, D.E. *et al.* (2003) Towards complete and accurate reporting of studies of diagnostic accuracy: the STARD initiative. *BMJ (Clinical Research Ed)*, **326** (7379), 41–44.

Boulesteix, A.L., Porzelius, C. and Daumer, M. (2008) Microarray-based classification and clinical predictors: on combined classifiers and additional predictive value. *Bioinformatics (Oxford, England)*, **24**, 1698–1706.

Boulesteix, A.L. and Strimmer, K. (2006) Partial least squares: a versatile tool for the analysis of high dimensional genomic data. *Briefings in Bioinformatics*, **8**, 32–44.

Boyle, J., Cavnor, C., Killcoyne, S. and Shmulevich, I. (2008) Systems biology driven software design for the research enterprise. *BMC Bioinformatics*, **9**, 295.

Braga-Neto, U.M. and Dougherty, E.R. (2004) Is cross-validation valid for small-sample microarray classification? *Bioinformatics (Oxford, England)*, **20**, 374–380.

Braunwald, E. (2008) Biomarkers in heart failure. *The New England Journal of Medicine*, **358**, 2148–2159.

Brazma, A., Krestyaninova, M. and Sarkans, U. (2006) Standards for systems biology. *Nature Reviews Genetics*, **7**, 593.

Brazma, A., Parkinson, H., Sarkans, U. *et al.* (2003) ArrayExpress–a public repository for microarray gene expression data at the EBI. *Nucleic Acids Research*, **31**, 68–71.

Breiman, L. (2001) Random forests. *Machine Learning*, **45**, 5–32.

Breiman, L., Friedman, J., Stone, C. and Olshen, R. (1984) *Classification and Regression Trees*, Chapman & Hall/CRC, Boca Raton, USA.

Brender, J. (2005) *Handbook of Evaluation Methods for Health Informatics*, Academic Press, Burlington, MA.

Brunet, J.P., Tamayo, P., Golub, T.R. and Mesirov, J.P. (2004) Metagenes and molecular pattern discovery using matrix factorization. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 4164–4169.

Bryan, K., Brennan, L. and Cunningham, P. (2008) MetaFIND: a feature analysis tool for metabolomics data. *BMC Bioinformatics*, **9**, 470.

Buckingham, S.D. (2008) Scientific software: seeing the SNPs between us. *Nature Methods*, **5**, 903–908.

Bull, T.M., Coldren, C.D., Moore, M. *et al.* (2004) Gene microarray analysis of peripheral blood cells in pulmonary arterial hypertension. *American Journal of Respiratory and Critical Care Medicine*, **170**, 911–919.

Bunch, T.J. and White, R.D. (2007) Decision making with biomarkers after cardiac arrest: are we there yet? *Critical Care Medicine*, **35** (5), 1411–1412.

Butte, A.J. (2008a) The ultimate model organism. *Science*, **320**, 325–327.

Butte, A.J. (2008b) Translational bioinformatics: coming of age. *Journal of the American Medical Informatics Association*, **15**, 709–714.

caBIG (2008) the cancer Biomedical Informatics Grid, [cabig.nci.nih.gov], accessed 18 December 2008.

Calvo, B., Larrañaga, P. and Lozano, J.A. (2007) Learning Bayesian classifiers from positive and unlabeled examples. *Pattern Recognition Letters*, **28**, 2375–2384.

Camargo, A. and Azuaje, F. (2007) Linking gene expression and functional network data in human heart failure. *PLoS One*, **2** (12), e1347.

Camargo, A. and Azuaje, F. (2008) Identification of dilated cardiomyopathy signature genes through gene expression and network data integration. *Genomics*, **92**, 404–413.

Cannataro, M. (2008) Computational proteomics: management and analysis of proteomics data. *Briefings in Bioinformatics*, **9**, 97–101.

Cannataro, M., Cuda, G., Gaspari, M. and Veltri, P. (2007) An interactive tool for the management and visualization of mass-spectrometry proteomics data, in *WILF-07, LNCS (LNAI)*, **4578** (eds F. Masulli, S. Mitra and G. Pasi), Springer, Heidelberg, pp. 635–642.

Carter, N.P. (2007) Methods and strategies for analyzing copy number variation using DNA microarrays. *Nature Genetics*, **39**, S16–S21.

Chen, D.P., Weber, S.C., Constantinou, P.S. *et al.* (2008a) Novel integration of hospital electronic medical records and gene expression measurements to identify genetic markers of maturation. *Pacific Symposium on Biocomputing*, **13**, 243–254.

Chen, H. and Sharp, B.M. (2004) Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics*, **5**, 147.

Chen, R., Li, L. and Butte, A.J. (2007) AILUN: reannotating gene expression data automatically. *Nature Methods*, **4**, 879.

Chen, R., Morgan, A.A., Dudley, J. *et al.* (2008b) FitSNPs: highly differentially expressed genes are more likely to have variants associated with disease. *Genome Biology*, **9**, R170.

Chen, T., Kao, M.Y., Tepel, M. *et al.* (2001) A dynamic programming approach to de novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **8**, 325–337.

Chittenden, T.W., Sherman, J.A., Xiong, F. *et al.* (2006) Transcriptional profiling in coronary artery disease: indications for novel markers of coronary collateralization. *Circulation*, **114**, 1811–1820.

Chuang, H.Y., Lee, E., Liu, Y.T. *et al.* (2007) Network-based classification of breast cancer metastasis. *Molecular Systems Biology*, **3**, 140.

Cleary, J.G. and Trigg, L. (1995) K*: an instance-based learner using an entropic distance measure. Proc Machine Learning '95, Lake Tahoe, USA, pp. 108–114.

Clifford, G.D., Azuaje, F. and McSharry, P.E. (eds) (2006) *Advanced Methods and Tools for ECG Analysis*, Artech House Publishing, London, UK.

Cline, M.S., Smoot, M., Cerami, E. *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape. *Nature Protocols*, **2**, 2366–2382.

Cobo, E., Selva-O'Callagham, A., Ribera, J.M. *et al.* (2007) Statistical reviewers improve reporting in biomedical articles: a randomized trial. *PLoS ONE*, **2** (3), e332.

Contopoulos-Ioannidis, D.G., Alexiou, G.A., Gouvias, T.C. and Ioannidis, J.P. (2008) Life cycle of translational research for medical interventions. *Science*, **321**, 1298–1299.

Cook, N.R. (2007) Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, **115**, 928–935.

Cook, N.R. (2008) Statistical evaluation of prognostic versus diagnostic models: beyond the ROC curve. *Clinical Chemistry*, **54**, 17–23.

Cordell, H.J. (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Human Molecular Genetics*, **11**, 2463–2468.

Couzin, J. (2008) Human genetics. Interest rises in DNA copy number variations–along with questions. *Science*, **322**, 1314.

Cover, T.M. and Thomas, J.A. (1991) *Elements of Information Theory*, Wiley.

Cox, D.G. and Kraft, P. (2006) Quantification of the power of Hardy-Weinberg equilibrium testing to detect genotyping error. *Human Heredity*, **61**, 10–14.

Crawford, S.L. (2006) Correlation and regression. *Circulation*, **114**, 2083–2088.

Csermely, P. (2006) *Weak Links: Stabilizers of Complex Systems from Proteins to Social Networks*, Springer, Berlin, Germany.

Curtis, D. (2007) Comparison of artificial neural network analysis with other multimarker methods for detecting genetic association. *BMC Genetics*, **8**, 49.

Daemen, A., Gevaert, O., Ojeda, F. *et al.* (2009) A kernel-based integration of genome-wide data for clinical decision support. *Genome Medicine*, **1**, 39.

Dancik, V., Addona, T.A., Clauser, K.R. *et al.* (1999) De novo peptide sequencing via tandem mass spectrometry. *Journal of Computational Biology: A Journal of Computational Molecular Cell Biology*, **6**, 327–342.

Davies, H.T., Crombie, I.K. and Tavakoli, M. (1998) When can odds ratios mislead? *BMJ (Clinical Research Ed)*, **316**, 989–991.

Davis, J. and Goadrich, M. (2006) The relationship between Precision-Recall and ROC curves. Proceedings of the 23rd international conference on Machine learning, Pittsburgh, Pennsylvania, pp. 233–240.

Davis, R.B. and Mukamal, K.J. (2006) Hypothesis testing: means. *Circulation*, **114** (10), 1078–1082.

De Bie, T., Tranchevent, L.C., van Oeffelen, L.M.M. and Moreau, Y. (2007) Kernel-based data fusion for gene prioritization. *Bioinformatics (Oxford, England)*, **23**, i125–i132.

De Jong, K.A. (2006) *Evolutionary Computation: A Unified Approach*, Bradford Book.

De Lemos, J.A. and Lloyd-Jones, D.M. (2008) Multiple biomarker panels for cardiovascular risk assessment. *The New England Journal of Medicine*, **358**, 2172–2174.

DePrimo, S.E., Wong, L.M., Khatry, D.B. *et al.* (2003) Expression profiling of blood samples from an SU5416 Phase III metastatic colorectal cancer clinical trial: a novel strategy for biomarker identification. *BMC Cancer*, **3**, 3.

Deschamps, A.M. and Spinale, F.G. (2006) Pathways of matrix metalloproteinase induction in heart failure: bioactive molecules and transcriptional regulation. *Cardiovascular Research*, **69**, 666–676.

Devarajan, K. (2008) Nonnegative matrix factorization: an analytical and interpretive tool in computational biology. *PLoS Computational Biology*, **4**, e1000029.

di Bernardo, D., Thompson, M.J., Gardner, T.S. *et al.* (2005) Chemogenomic profiling on a genome-wide scale using reverseengineered gene networks. *Nature Biotechnology*, **23**, 377–383.

Díaz-Uriarte, R. and Alvarez de Andrés, S. (2006) Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, **7**, 3.

Dinov, I.D., Rubin, D., Lorensen, W. *et al.* (2008) iTools: a framework for classification, categorization and integration of computational biology resources. *PLoS ONE*, **3**, e2265.

Dobbin, K. and Simon, R. (2005) Sample size determination in microarray experiments for class comparison and prognostic classification. *Biostatistics*, **6**, 27–38.

Domon, B. and Aebersold, R. (2006) Mass spectrometry and protein analysis. *Science*, **312**, 212–217.

Donaldson, I., Martin, J., de Bruijn, B. *et al.* (2003) PreBIND and Textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC Bioinformatics*, **4**, 11.

Dong, J. and Horvath, S. (2007) Understanding network concepts in modules. *BMC Systems Biology*, **1**, 24.

Donnelly, P. (2008) Progress and challenges in genome-wide association studies in humans. *Nature*, **456**, 728–731.

Dos Remedios, C.G., Liew, C.C., Allen, P.D. *et al.* (2003) Genomics, proteomics and bioinformatics of human heart failure. *Journal of Muscle Research and Cell Motility*, **24**, 251–260.

Drake, T.A. and Ping, P. (2007) Thematic review series: systems biology approaches to metabolic and cardiovascular disorders. Proteomics approaches to the systems biology of cardiovascular diseases. *Journal of Lipid Research*, **48**, 1–8.

Dudley, J.T. and Butte, A.J. (2009) Identification of discriminating biomarkers for human disease using integrative network biology. *Pacific Symposium on Biocomputing*, 27–38.

Dudoit, S., Fridlyand, J. and Speed, T.P. (2002) Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, **97**, 77–87.

Dudoit, S. and van der Laan, M.J. (2008) *Multiple Testing Procedures with Applications to Genomics*, Springer, New York, USA.

Duncan, M.W. and Hunsucker, S.W. (2005) Proteomics as a tool for clinically relevant biomarker discovery and validation. *Experimental Biology and Medicine*, **230**, 808–817.

Dupont, W.D. and Plummer, W.D. (2008) PS: Power and Sample Size Calculation, version 2.1.31, 2004, http://biostat.mc.vanderbilt.edu/twiki/bin/view/Main/PowerSampleSize.

Dupuy, A. and Simon, R.M. (2007) Critical review of published microarray studies for cancer outcome and guidelines on statistical analysis and reporting. *Journal of the National Cancer Institute*, **99**, 147–157.

Dwan, K., Altman, D.G., Arnaiz, J.A. *et al.* (2008) Systematic review of the empirical evidence of study publication bias and outcome reporting bias. *PLoS ONE*, **3** (8), e3081.

Eady, J.J., Wortley, G.M., Wormstone, Y.M. *et al.* (2005) Variation in gene expression profiles of peripheral blood mononuclear cells from healthy volunteers. *Physiological Genomics*, **22**, 402–411.

Eddy, S.R. (2005) 'Antedisciplinary' science. *PLoS Computational Biology*, **1**, e6.

EDRN (2008) The NCI's Early Detection Research Network, [edrn.nci.nih.gov], 22 December 2008.

Efron, B. and Tibshirani, R.J. (1997) Improvements on cross-validation: the .632 + bootstrap method. *Journal of the American Statistical Association*, **92**, 548–560.

Efron, B. and Tibshirani, R. (1993) *An Introduction to Bootstrap*, Chapman and Hall, New York.

Efroni, S., Schaefer, C.F. and Buetow, K.H. (2007) Identification of key processes underlying cancer phenotypes using biologic pathway analysis. *PLoS ONE*, **2**, e425.

Ein-Dor, L., Kela, I., Getz, G. *et al.* (2005) Outcome signature genes in breast cancer: is there a unique set? *Bioinformatics (Oxford, England)*, **21**, 171–178.

El-Omar, E.M., Ng, M.T. and Hold, G.L. (2008) Polymorphisms in Toll-like receptor genes and risk of cancer. *Oncogene*, **27**, 244–252.

Eng, J. (2003) Sample size estimation: how many individuals should be studied. *Radiology*, **227**, 309–313.

Eng, J. (2006) ROC analysis: web-based calculator for ROC curves. Baltimore: Johns Hopkins University. Available from: http://www.jrocfit.org.

English, S.B. and Butte, A.J. (2007) Evaluation and integration of 49 genomewide experiments and the prediction of previously unknown obesity-related genes. *Bioinformatics (Oxford, England)*, **23**, 2910–2917.

Ergün, A., Lawrence, C.A., Kohanski, M.A. *et al.* (2007) A network biology approach to prostate cancer. *Molecular Systems Biology*, **3**, 82.

Ernens, I., Rouy, D., Velot, E. *et al.* (2006) Adenosine inhibits matrix metalloproteinase-9 secretion by neutrophils: implication of A2a receptor and cAMP/PKA/Ca2 + pathway. *Circulation Research*, **99** (6), 590–597.

Ewens, W.J. and Grant, G.R. (2005) *Statistical Methods in Bioinformatics*, 2nd edn, Springer, NY.

Fan, C., Oh, D.S., Wessels, L. *et al.* (2006) Concordance among gene-expression-based predictors for breast cancer. *The New England Journal of Medicine*, **355**, 560–569.

Feezor, R.J., Baker, H.V., Mindrinos, M. *et al.* (2004) Whole blood and leukocyte RNA isolation for gene expression analyses. *Physiological Genomics*, **19**, 247–254.

Feller, J., Fitzgerald, B., Hissam, S.A. and Lakhani, K.R. (eds) (2005) *Perspectives on Free and Open Source Software*, MIT Press, Cambridge, MA, USA.

Fernández-Suárez, X.M. and Birney, E. (2008) Advanced genomic data mining. *PLoS Computational Biology*, **4** (9), e1000121.

Fogel, G.B., Corne, D.W. and Pan, Y. (eds) (2008) *Computational Intelligence in Bioinformatics*, WileyBlackwell, Hoboken, NJ, USA.

Frank, A.M., Bandeira, N., Shen, Z. *et al.* (2008) Clustering millions of tandem mass spectra. *Journal of Proteome Research*, **7**, 113–122.

Frank, E., Hall, M., Trigg, L. *et al.* (2004) Data mining in bioinformatics using Weka. *Bioinformatics (Oxford, England)*, **20**, 2479–2481.

Frankel, D.S., Piette, J.D., Jessup, M. *et al.* (2006) Validation of prognostic models among patients with advanced heart failure. *Journal of Cardiac Failure*, **12**, 430–438.

Freeman, J.L., Perry, G.H., Feuk, L. *et al.* (2006) Copy number variation: new insights in genome diversity. *Genome Research*, **16**, 949–961.

Freeman, T.C., Goldovsky, L., Brosch, M. *et al.* (2007) Construction, visualisation, and clustering of transcription networks from microarray expression data. *PLoS Computational Biology*, **3**, 2032–2042.

Freimer, N. and Sabatti, C. (2003) The human phenome project. *Nature Genetics*, **34**, 15–21.

Frey, B.J. and Dueck, D. (2007) Clustering by passing messages between data points. *Science*, **315**, 972–976.

Gadbury, G.L. Page, G.P., Edwards, J., Kayo, T., Weindruch, R., Permana, P.A., Mountz, J. and Allison, D.B. (2004) Power analysis and sample size estimation in the age of high dimensional biology. *Stat Meth Med Res*, **13**, 325–338.

Ganesh, S.K., Sharma, Y., Dayhoff, J. *et al.* (2007) Detection of venous thromboembolism by proteomic serum biomarkers. *PLoS ONE*, **2**, e544. doi: 10.1371/journal.pone.0000544.

Gao, X., Starmer, J. and Martin, E.R. (2008) A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic Epidemiology*, **32**, 361–369.

Gargalovic, P.S., Imura, M., Zhang, B. *et al.* (2006) Identification of inflammatory gene modules based on variations of human endothelial cell responses to oxidized lipids. *Proceedings of the National Academy of Sciences of the United States of America*, **103**, 12741–12746.

Gauderman, W.J., Murcray, C., Gilliland, F. and Conti, D.V. (2007) Testing association between disease and multiple SNPs in a candidate gene. *Genetic Epidemiology*, **31**, 383–395.

Gauvreau, K. (2006) Hypothesis testing: proportions. *Circulation*, **114** (14), 1545.

Gentleman, R., Carey, V., Huber, W. *et al.* (eds) (2005) *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Springer, NY, USA.

Gentleman, R.C., Carey, V.J., Bates, D.M. *et al.* (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biology*, **5** (10), R80.

GEO, Gene Expression Omnibus (2008) http://www.ncbi.nlm.nih.gov/geo/, last accessed 10 April 2009.

Gerszten, R.E. and Wang, T.J. (2008) The search for new cardiovascular biomarkers. *Nature*, **451**, 949–952.

Gewin, V. (2007) Missing the mark. *Nature*, **449**, 770–771.

geWorkbench (2008) [wiki.c2b2.columbia.edu/workbench], 22 December 2008.

Ghazalpour, A., Doss, S., Zhang, B. *et al.* (2006) Integrating genetic and network analysis to characterize genes related to mouse weight. *PLoS Genetics*, **2**, e130.

Ghosh, D. and Poisson, L.M. (2009) 'Omics' data and levels of evidence for biomarker discovery. *Genomics*, **93**, 13–16.

Ghosh, S., Grant, D.F., Dey, D.K. and Hill, D.W. (2008) A semiparametric modeling framework for potential biomarker discovery and the development of metabonomic profiles. *BMC Bioinformatics*, **9**, 38.

Ginsburg, G.S., Seo, D. and Frazier, C. (2006) Microarrays coming of age in cardiovascular medicine: standards, predictions, and biology. *Journal of the American College of Cardiology*, **48**, 1618–1620.

Glantz, S.A. (2001) *Primer of Biostatistics*, 5th edn, McGraw-Hill/Appleton & Lange.

Glantz, S.A. and Slinker, B.K. (2001) *Primer of Applied Regression and Analysis of Variance*, 2nd edn, McGraw-Hill.

Goeman, J.J. and Bühlmann, P. (2007) Analyzing gene expression data in terms of gene sets: methodological issues. *Bioinformatics (Oxford, England)*, **23**, 980–987.

Goh, K.I., Cusick, M.E., Valle, D. *et al.* (2007) The human disease network. *Proceedings of the National Academy of Sciences of the United States of America*, **104**, 8685–8690.

Golden Helix (2009) viewed 10 January 2009, http://www.goldenhelix.com/.

Good, P.I. (2006) *Resampling Methods*, 3rd edn, Birkhäuser, Boston.

Grzeskowiak, R., Witt, H., Drungowski, M. *et al.* (2003) Expression profiling of human idiopathic dilated cardiomyopathy. *Cardiovascular Research*, **59**, 400–411.

Guo, Z., Zhang, T., Li, X. *et al.* (2005) Towards precise classification of cancers based on robust gene functional expression profiles. *BMC Bioinformatics*, **6**, 58.

Guyon, I. and Elisseeff, A. (2003) An introduction to variable and feature selection. *Journal of Machine Learning Research*, **3**, 1157–1182.

Haider, N. (2008) http://merian.pch.univie.ac.at/∼nhaider/cheminf/jdxview.html, last accessed 30 March 2009.

Hakes, L., Pinney, J.W., Robertson, D.L. and Lovell, S.C. (2008) Protein-protein interaction networks and biology–what's the connection? *Nature Biotechnology*, **26**, 69–72.

Hall, M.A. (1999) Correlation-based Feature Subset Selection for Machine Learning. PhD thesis, Department of Computer Science, University of Waikato.

Hanash, S.M., Pitteri, S.J. and Faca, V.M. (2008) Mining the plasma proteome for cancer biomarkers. *Nature*, **452**, 571–579.

Hanley, J.A. and McNeil, B.J. (1982) The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36.

Hanley, J.A. and McNeil, B.J. (1983) A method of comparing the areas under receiver operating characteristic curves derived from the same cases. *Radiology*, **148**, 839–843.

Hardy, J. and Singleton, A. (2008) The HapMap: charting a course for genetic discovery in neurological diseases. *Archives of Neurology*, **65**, 319–321.

Hastie, T., Tibshirani, R. and Friedman, J.H. (2001) *The Elements of Statistical Learning*, Springer, NY, USA.

Hauskrecht, M., Pelikan, R., Valko, M. and Lyons-Weiler, J. (2006) Feature selection and dimensionality reduction in genomics and proteomics, in *Fundamentals of Data Mining in Genomics and Proteomics* (eds D.P. Berrar, W. Dubitzky and M. Granzow), Springer, pp. 149–172.

Heidecker, B., Kasper, E.K., Wittstein, I.S. *et al.* (2008) Transcriptomic biomarkers for individual risk assessment in new-onset hear failure. *Circulation*, **118**, 238–246.

Henney, A. and Superti-Furga, G. (2008) A network solution. *Nature*, **455**, 730–731.

Heredia-Langner, A., Cannon, W.R., Jarman, K.D. and Jarman, K.H. (2004) Sequence optimization as an alternative to de novo analysis of tandem mass spectrometry data. *Bioinformatics (Oxford, England)*, **20**, 2296–2304.

Hewett, M., Oliver, D.E., Rubin, D.L. *et al.* (2002) PharmGKB: the pharmacogenetics knowledge base. *Nucleic Acids Research*, **30**, 163–165.

Higdon, R., van Belle, G. and Kolker, E. (2008) A note on the false discovery rate and inconsistent comparisons between experiments. *Bioinformatics (Oxford, England)*, **24**, 1225–1228.

Hilario, M. and Kalousis, A. (2008) Approaches to dimensionality reduction in proteomic biomarker studies. *Briefings in Bioinformatics*, **9**, 102–118.

Hilario, M., Kalousis, A., Müller, M. and Pellegrini, C. (2003) Machine learning approaches to lung cancer prediction from mass spectra. *Proteomics*, **3**, 1716–1719.

Hindorff, L.A., Junkins, H.A., Mehta, J.P. and Manolio, T.A. (2008) A Catalog of Published Genome-Wide Association Studies. www.genome.gov/26525384, Accessed 11 January 2009.

Hinestrosa, M.C., Dickersin, K., Klein, P. *et al.* (2007) Shaping the future of biomarker research in breast cancer to ensure clinical relevance. *Nature Reviews. Cancer*, **7**, 309–315.

Hirschhorn, J.N. and Daly, M.J. (2005) Genome-wide association studies for common diseases and complex traits. *Nature Reviews. Genetics*, **6**, 95–108.

Hoffmann, R. and Valencia, A. (2005) Implementing the iHOP concept for navigation of biomedical literature. *Bioinformatics (Oxford, England)*, **21** (Suppl. 2), ii252–ii258.

Holmes, E., Nicholls, A.W., Lindon, J.C. *et al.* Chemometric models for toxicity classification based on NMR spectra of biofluids. *Chemical Research in Toxicology*, **13** (6), 471–478.

Hosking, L., Lumsden, S., Lewis, K. *et al.* (2004) Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *European Journal of Human Genetics*, **12**, 395–399.

Hua, J., Xiong, Z., Lowey, J. *et al.* (2005) Optimal number of features as a function of sample size for various classification rules. *Bioinformatics (Oxford, England)*, **21**, 1509–1515.

Huang, T., Tu, K., Shyr, Y. *et al.* (2008) The prediction of interferon treatment effects based on time series microarray gene expression profiles. *Journal of Translational Medicine*, **6**, 44.

Hull, D., Wolstencroft, K., Stevens, R. *et al.* (2006) Taverna: a tool for building and running workflows of services. *Nucleic Acids Research*, **34**, 729–732.

Ideker, T. and Sharan, R. (2008) Protein networks in disease. *Genome Research*, **18** (4), 644–652.

Ihmels, J., Friedlander, G., Bergmann, S. *et al.* (2002) Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, **31**, 370–377.

Ioannidis, J.P. (2006) Commentary: grading the credibility of molecular evidence for complex diseases. *International Journal of Epidemiology*, **35**, 572–578.

Ioannidis, J.P. (2007a) Limitations are not properly acknowledged in the scientific literature. *Journal of Clinical Epidemiology*, **60**, 324–329.

Ioannidis, J.P. (2007b) Is molecular profiling ready for use in clinical decision making? *Oncologist*, **12**, 301–311.

Ioannidis, J.P., Allison, D.B., Ball, C.A. *et al.* (2009) Repeatability of published microarray gene expression analyses. *Nature Genetics*, **41**, 149–155.

IPA (2009) Ingenuity Pathway Analysis (IPA), http://www.ingenuity.com/ last time accessed: 3 May 2009.

Jafari, P. and Azuaje, F. (2006) An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, **6**, 27.

Jansen, J.J., Hoefsloot, H.C., Boelens, H.F. *et al.* (2004) Analysis of longitudinal metabolomics data. *Bioinformatics (Oxford, England)*, **20**, 2438–2446.

Jensen, F.V. and Nielsen, T. (2007) *Bayesian Networks and Decision Graphs*, Springer-Verlag, New York.

Jiao, S. and Zhang, S. (2008) On correcting the overestimation of the permutation-based false discovery rate estimator. *Bioinformatics (Oxford, England)*, **24**, 1655–1661.

Jonsson, P.F. and Bates, P.A. (2006) Global topological features of cancer proteins in the human interactome. *Bioinformatics (Oxford, England)*, **22**, 2291–2297.

JPF (2008) [jpf.sourceforge.net], last accessed 21 May 2008.

Kalorama Information (2007) Biomarkers: A Market Briefing, [www.kaloramainformation.com].

Karolchik, D., Kuhn, R.M., Baertsch, R. *et al.* (2008) The UCSC Genome Browser Database: 2008 update. *Nucleic Acids Research*, **36** (Database issue), D773–D779.

Kardys, I., Knetsch, A.M., Bleumink, G.S. *et al.* (2006) C-reactive protein and risk of heart failure. The Rotterdam Study. 2006. *American Heart Journal*, **152**, 514–520.

Karlebach, G. and Shamir, R. (2008) Modelling and analysis of gene regulatory networks. *Nature Reviews. Molecular Cell Biology*, **9**, 770–780.

Katajamaa, M., Miettinen, J. and Orešič, M. (2006) MZmine: toolbox for processing and visualization of mass spectrometry based molecular profile data. *Bioinformatics (Oxford, England)*, **22**, 634–636.

Kathiresan, S., Melander, O., Guiducci, C. *et al.* (2008) Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genetics*, **40**, 189–197.

Kawasaki, M., Iwasaki, M., Koshiba, T. *et al.* (2007) Gene expression profile analysis of the peripheral blood mononuclear cells from tolerant living-donor liver transplant recipients. *International surgery*, **92** (5), 276–286.

Keator, D.B., Grethe, J.S., Marcus, D. *et al.* (2008) A national human neuroimaging collaboratory enabled by the Biomedical Informatics Research Network (BIRN). *IEEE Transactions on Information Technology in Biomedicine*, **12**, 162–172.

Keller, M.P., Choi, Y., Wang, P. *et al.* (2008) A gene expression network model of type 2 diabetes links cell cycle regulation in islets with diabetes susceptibility. *Genome Research*, **18**, 706–716.

Kingsford, C. and Salzberg, S.L. (2008) What are decision trees? *Nature Biotechnology*, **26** (9), 1011–1013.

Kitano, H. (2002) Computational systems biology. *Nature*, **420** (6912), 206–210.

Kittler, J. (2000) A framework for classifier fusion: is it still needed? Proceedings of the Joint IAPR International Workshops on Advances in Pattern Recognition, Lecture Notes In Computer Science, 1876, pp. 45–56.

Kittler, J., Hatef, M., Duin, R.P.W. and Matas, J. (1998) On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**, 226–239.

Kittleson, M.M. and Hare, J.M. (2005) Molecular signature analysis: using the myocardial transcriptome as a biomarker in cardiovascular disease. *Trends in Cardiovascular Medicine*, **15**, 130–138.

Kittleson, M.M., Ye, S.Q., Irizarry, R.A. *et al.* (2004) Identification of a gene expression profile that differentiates between ischemic and nonischemic cardiomyopathy. *Circulation*, **110**, 3444–3451.

Klein, J.P. and Moeschberger, M.L. (2005) *Survival Analysis*, Springer, NY.

Klein, R.J. (2007) Power analysis for genome-wide association studies. *BMC Genetics*, **8**, 58.

Kleinbaum, D. and Klein, M. (2005a) *Survival Analysis: A Self-Learning Text*, 2nd edn, Springer, NY.

Kleinbaum, D.G. and Klein, M. (2005b) *Logistic Regression*, 2nd edn, Springer, NY, USA.

Knickerbocker, T., Chen, J.R., Thadhani, R. and MacBeath, G. (2007) An integrated approach to prognosis using protein microarrays and nonparametric methods. *Molecular Systems Biology*, **3**, 123.

Kohavi, R. and John, G. (1997) Wrappers for feature selection. *Artificial Intelligence*, **97**, 273–324.

Kohonen, T. (2000) *Self-Organizing Maps*, 3rd edn, Springer.

König, I.R., Malley, J.D., Pajevic, S. *et al.* (2008) Patient-centered yes/no prognosis using learning machines. *International Journal of Data Mining and Bioinformatics*, **2**, 289–341.

Kopka, J., Schauer, N., Krueger, S. *et al.* (2005) GMD@CSB.DB: the golm metabolome database. *Bioinformatics (Oxford, England)*, **21**, 1635–1638.

Kostka, D. and Spang, R. (2008) Microarray based diagnosis profits from better documentation of gene expression signatures. *PLoS Computational Biology*, **4** (2), e22.

Kruglyak, L. (2008) The road to genome-wide association studies. *Nature Reviews. Genetics*, **9**, 314–318.

Lage, K., Karlberg, E.O., Storling, Z.M. *et al.* (2007) A human phenome-interactome network of protein complexes implicated in genetic disorders. *Nature Biotechnology*, **25**, 309–316.

Lanckriet, G.R.G., De Bie, T., Cristianini, N. *et al.* (2004) A statistical framework for genomic data fusion. *Bioinformatics (Oxford, England)*, **20**, 2626–2635.

Landegren, U., Nilsson, M. and Kwok, P.Y. (1998) Reading bits of genetic information: methods for single-nucleotide polymorphism analysis. *Genome Research*, **8**, 769–776.

Larson, M.G. (2006) Descriptive statistics and graphical displays. *Circulation*, **114** (1), 76–81.

LeadDiscovery, 1 March 2008, "Biomarkers-technologies, markets and companies", [www.leaddiscovery.co.uk].

Ledford, H. (2008) Drug markers questioned. *Nature*, **452**, 510–511.

Lee, E., Chuang, H.-Y., Kim, J.-W. *et al.* (2008) Inferring pathway activity toward precise disease classification. *PLoS Computational Biology*, **4** (11), e1000217.

Lee, J.W., Lee, J.B., Park, M. and Song, S.H. (2005) An extensive evaluation of recent classification tools applied to microarray data. *Computation Statistics and Data Analysis*, **48**, 869–885.

Lefebvre, C., Lim, W.K., Basso, K. *et al.* (2007) A context-specific network of protein–DNA and protein–protein interactions reveals new regulatory motifs in human B cells. *Lecture Notes in Bioinformatics*, **4532**, 42–56.

Lenth, R.V. (2000) Java Applets for Power and Sample Size, http://www.stat.uiowa.edu/~rlenth/Power/.

Lenth, R.V. (2001) Some practical guidelines for effective sample size determination. *The American Statistician*, **55**, 187–193.

Lewis, G.D., Asnani, A. and Gerszten, R.E. (2008) Application of metabolomics to cardiovascular biomarker and pathway discovery. *Journal of the American College of Cardiology*, **52**, 117–123.

Li, J. and Li, L. (2005) Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, **95**, 221–227.

Li, T., Zhang, C. and Ogihara, M. (2004a) A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics*, **20**, 2429–2437.

Li, W. (2008) Three lectures on case-control genetic association analysis. *Briefings in Bioinformatics*, **9**, 1–13.

Li, X.J., Pedrioli, P.G., Eng, J. *et al.* (2004b) A tool to visualize and evaluate data obtained by liquid chromatography-electrospray ionization-mass spectrometry. *Analytical Chemistry*, **76**, 3856–3860.

Liebman, M.N. (2005) An engineering approach to translational medicine. *American Scientist*, **93**, 296–298.

Liew, C.C. (2005) Expressed genome molecular signatures of heart failure. *Clinical Chemistry and Laboratory Medicine: CCLM/FESCC*, **43**, 462–469.

Liew, C.C., Ma, J., Tang, H.C. *et al.* (2006) The peripheral blood transcriptome dynamically reflects system wide biology: a potential diagnostic tool. *The Journal of Laboratory and Clinical Medicine*, **147**, 126–132.

Lim, W.K., Lyashenko, E. and Califano, A. (2009) Master regulators used as breast cancer metastasis classifier. *Pacific Symposium on Biocomputing*, **14**, 504–515.

Lindholm, E., Melander, O., Almgren, P. *et al.* (2006) Polymorphism in the MHC2TA gene is associated with features of the metabolic syndrome and cardiovascular mortality. *PLoS ONE*, **1**, e64.

Loscalzo, J. (2007) Association studies in an era of too much information: clinical analysis of new biomarker and genetic data. *Circulation*, **116**, 1866–1870.

Loscalzo, J. (2009) Pilot trials in clinical research: of what value are they? *Circulation*, **19** (13), 1694–1696.

Loscalzo, J., Kohane, I. and Barabasi, A.L. (2007) Human disease classification in the postgenomic era: a complex systems approach to human pathobiology. *Molecular Systems Biology*, **3**, 124.

Lu, X., Jain, V.V., Finn, P.W. and Perkins, D.L. (2007) Hubs in biological interaction networks exhibit low changes in expression in experimental asthma. *Molecular Systems Biology*, **3**, 98.

Lupski, J.R. (2007) Structural variation in the human genome. *The New England Journal of Medicine*, **356**, 1169–1171.

Luscombe, N.M., Babu, M.M., Yu, H. *et al.* (2004) Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature*, **431**, 308–312.

Lutz, A.M., Willmann, J.K., Cochran, F.V. *et al.* (2008) Cancer screening: a mathematical model relating secreted blood biomarker levels to tumor sizes. *PLoS Medicine*, **5**, e170.

Ma, S. and Huang, J. (2005) Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics (Oxford, England)*, **21**, 4356–4362.

Ma, J. and Liew, C.C. (2003) Gene profiling identifies secreted protein transcripts from peripheral blood cells in coronary artery disease. *Journal of Molecular and Cellular Cardiology*, **35**, 993–998.

Mailman, M.D., Feolo, M., Jin, Y. *et al.* (2007) The NCBI dbGaP database of genotypes and phenotypes. *Nature Genetics*, **39**, 1181–1186.

Maisel, A. (2007) Biomarkers in heart failure. Does prognostic utility translate to clinical futility? *Journal of the American College of Cardiology*, **50** (11), 1061–1063.

Malovini, A., Nuzzo, A., Ferrazzi, F. *et al.* (2009) Phenotype forecasting with SNPs data through gene-based Bayesian networks. *BMC Bioinformatics*, **10** (Suppl 2), S7.

Mamtani, M.R., Thakre, T.P., Kalkonde, M.Y. *et al.* (2006) A simple method to combine multiple molecular biomarkers for dichotomous diagnostic classification. *BMC Bioinformatics*, **7**, 442.

Mani, K.M., Lefebvre, C., Wang, K. *et al.* (2008) A systems biology approach to prediction of oncogenes and molecular perturbation targets in B-cell lymphomas. *Molecular Systems Biology*, **4**, 16.

Marcus, F.B. (2008) *Bioinformatics and Systems Biology: Collaborative Research and Resources*, Springer, Berlin, Germany.

Margolin, A.A., Wang, K., Lim, W.K. *et al.* (2006) Reverse engineering cellular networks. *Nature Protocols*, **1**, 662–671.

Markley, J.L., Anderson, M.E., Cui, Q. *et al.* (2007) New bioinformatics resources for metabolomics. *Pacific Symposium on Biocomputing*, **12**, 157–168.

Martens, L., Hermjakob, H., Jones, P. *et al.* (2005) PRIDE: the proteomics identifications database. *Proteomics*, **5**, 3537–3545.

Mathew, J.P., Taylor, B.S., Bader, G.D. *et al.* (2007) From bytes to bedside: data integration and computational biology for translational cancer research. *PLoS Computational Biology*, **3** (2), e12.

Mayr, M. (2008) Metabolomics: ready for the prime time? *Circulation: Cardiovascular Genetics*, **1**, 58–65.

McShane, L.M., Altman, D.G., Sauerbrei, W. *et al.* (2005) REporting recommendations for tumour MARKer prognostic studies (REMARK). *British Journal of Cancer*, **93**, 387–391.

Mechanic, L.E., Luke, B.T., Goodman, J.E. *et al.* (2008) Polymorphism Interaction Analysis (PIA): a method for investigating complex gene-gene interactions. *BMC Bioinformatics*, **9**, 146.

Mehra, M.R., Kobashigawa, J.A., Deng, M.C. *et al.* (2008) Clinical implications and longitudinal alteration of peripheral blood transcriptional signals indicative of future cardiac allograft rejection. *The Journal of Heart and Lung Transplantation*, **27**, 297–301.

Mejía-Roa, E., Carmona-Saez, P., Nogales, R. *et al.* (2008) bioNMF: a web-based tool for nonnegative matrix factorization in biology. *Nucleic Acids Research*, **36**, W523–W528.

Mendes, P. (2002) Emerging bioinformatics for the metabolome. *Briefings in Bioinformatics*, **3**, 134–145.

Milne, R.L., Ribas, G., González-Neira, A. *et al.* (2006) ERCC4 associated with breast cancer risk: a two-stage case-control study using high-throughput genotyping. *Cancer Research*, **66**, 9420–9427.

Mjolsness, E. and DeCoste, D. (2001) Machine learning for science: state of the art and future prospects. *Science*, **293**, 2051–2055.

Moher, D., Cook, D.J., Eastwood, S. *et al.* (1999) Improving the quality of reports of meta-analyses of randomised controlled trials: the QUOROM statement. Quality of Reporting of Meta-analyses. *Lancet*, **354**, 1896–1900.

Molinaro, A.M., Simon, R. and Pfeiffer, R.M. (2005) Prediction error estimation: a comparison of resampling methods. *Bioinformatics (Oxford, England)*, **21**, 3301–3307.

Montaner, D., Tárraga, J., Huerta-Cepas, J. *et al.* (2006) Next station in microarray data analysis: GEPAS. *Nucleic Acids Research*, **34** (Web Server issue), W486–W491.

Montero-Conde, C., Martín-Campos, J.M., Lerma, E. *et al.* (2008) Molecular profiling related to poor prognosis in thyroid carcinoma. Combining gene expression data and biological information. *Oncogene*, **27** (11), 1554–1561.

Moore, D.F., Li, H., Jeffries, N. *et al.* (2005) Using peripheral blood mononuclear cells to determine a gene expression profile of acute ischemic stroke: a pilot investigation. *Circulation*, **111**, 212–221.

Musunuru, K. and Kathiresan, S. (2008) HapMap and mapping genes for cardiovascular disease. *Circulation: Cardiovascular Genetics*, **1**, 66–71.

Nanni, L., Romualdi, C., Maseri, A. and Lanfranchi, G. (2006) Differential gene expression profiling in genetic and multifactorial cardiovascular diseases. *Journal of Molecular and Cellular Cardiology*, **41**, 934–948.

NCBCS (2009) The U.S. National Centers for Biomedical Computing, [http://www.ncbcs.org/], last accessed: 12 April 2009.

NCI-NHGRI Working Group on Replication in Association Studies. (2007) Replicating genotype-phenotype associations. *Nature*, **447**, 655–660.

Needham, C.J., Bradford, J.R., Bulpitt, A.J. and Westhead, D.R. (2006) Inference in Bayesian networks. *Nature Biotechnology*, **24**, 51–53.

Needham, C.J., Bradford, J.R., Bulpitt, A.J. and Westhead, D.R. (2007) A primer on learning in Bayesian networks for computational biology. *PLoS Computational Biology*, **3**, e129.

Nian, M., Lee, P., Khaper, N. and Liu, P. (2004) Inflammatory cytokines and postmyocardial infarction remodeling. *Circulation Research*, **94**, 1543–1553.

Nicholson, J.K. and Lindon, J.C. (2008) Systems biology: metabonomics. *Nature*, **455**, 1054–1056.

Nielsen, D.M., Ehm, M.G. and Weir, B.S. (1998) Detecting marker-disease association by testing for Hardy-Weinberg disequilibrium at a marker locus. *American Journal of Human Genetics*, **63**, 1531–1540.

Noble, D. (2006b) *The Music of Life: Biology beyond the Genome*, OUP, Oxford, UK.

Noble, W.S. (2006a) What is a support vector machine? *Nature Biotechnology*, **24**, 1565–1567.

Nolan, J.P., Deakin, C.D., Soar, J. *et al.* (2005) European Resuscitation Council Guidelines for Resuscitation 2005 Section 4. Adult advanced life support. *Resuscitation*, **67S1**, 39–86.

Nurse, P. (2008) Life, logic and information. *Nature*, **454**, 424–426.

Nyholt, D.R. (2004) A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *American Journal of Human Genetics*, **74**, 765–769.

Oh, J.H., Kim, Y.B., Gurnani, P. *et al.* (2008) Biomarker selection and sample prediction for multi-category disease on MALDI-TOF data. *Bioinformatics (Oxford, England)*, **24**, 1812–1818.

Oinn, T., Addis, M., Ferris, J. *et al.* (2004) Taverna: a tool for the composition and enactment of bioinformatics workflows. *Bioinformatics (Oxford, England)*, **20**, 3045–3054.

Omenn, G.S., States, D.J., Adamski, M. *et al.* (2005) Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 3020 proteins and a publicly-available database. *Proteomics*, **5**, 3226–3245.

OMIM (2009) Online Mendelian Inheritance in Man. http://www.ncbi.nlm.nih.gov/omim/ Last accessed: 28 April 2009.

Oti, M., Snel, B., Huynen, M.A. and Brunner, H.G. (2006) Predicting disease genes using protein-protein interactions. *Journal of Medical Genetics*, **43**, 691–698.

Page, G.P., Edwards, J.W., Gadbury, G.L. *et al.* (2006) The PowerAtlas: a power and sample size atlas for microarray experimental design and research. *BMC Bioinformatics*, **7**, 84.

Palagi, P.M., Walther, D., Quadroni, M. *et al.* (2005) MSight: an image analysis software for liquid chromatography-mass spectrometry. *Proteomics*, **5**, 2381–2384.

Papadopoulos, M.C., Abel, P.M., Agranoff, D. *et al.* (2004) A novel and accurate diagnostic test for human African trypanosomiasis. *Lancet*, **363**, 1358–1363.

Parsons, H.M., Ludwig, C., Günther, U.L. and Viant, M.R. (2007) Improved classification accuracy in 1- and 2-dimensional NMR metabolomics data using the variance stabilising generalised logarithm transformation. *BMC Bioinformatics*, **8**, 234.

Pascual-Montano, A., Carmona-Saez, P., Chagoyen, M. *et al.* (2006) bioNMF: a versatile tool for non-negative matrix factorization in biology. *BMC Bioinformatics*, **7**, 366.

Patino, W.D., Mian, O.Y., Kang, J.G. *et al.* (2005) Circulating transcriptome reveals markers of atherosclerosis. *Proceedings of the National Academy of Sciences of the United States of America*, **102** (9), 3423–3428.

Peacock, W.F. 4th, De Marco, T., Fonarow, G.C. *et al.* (2008) Cardiac troponin and outcome in acute heart failure. *The New England Journal of Medicine*, **358** (20), 2117–2126.

Pearson, H. (2007) Meet the human metabolome. *Nature*, **446**, 8.

Pencina, M.J., D'Agostino, R.B. Sr, D'Agostino, R.B. Jr and Vasan, R.S. (2008) Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in Medicine*, **27**, 157–172.

Pepe, M.S. (2008) webpage at the Fred Hutchinson Cancer Research Center, https://www.fhcrc.org/science/phs/biostats/pepe.html.

Pepe, M.S., Cai, T. and Longton, G. (2006) Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics*, **62**, 221–229.

Pepe, M.S., Etzioni, R., Feng, Z. *et al.* (2001) Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute*, **93**, 1054–1061.

Pepe, M.S., Feng, Z. and Gu, J.W. (2007) Comments on 'Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond' by M. J. Pencina *et al. Statistics in Medicine*, **27**, 173–181.

Pepe, M.S., Feng, Z., Huang, Y. *et al.* (2008a) Integrating the predictiveness of a marker with its performance as a classifier. *American Journal of Epidemiology*, **167**, 362–368.

Pepe, M.S., Janes, H. and Gu, J.W. (2007) Letter by Pepe *et al.* regarding article, "Use and misuse of the receiver operating characteristic curve in risk prediction". *Circulation*, **116** (6), e132.

Pepe, M.S. and Janes, H.E. (2008) Gauging the performance of SNPs, biomarkers, and clinical factors for predicting risk of breast cancer. *Journal of the National Cancer Institute*, **100**, 978–979.

Pepe, M.S., Zheng, Y., Jin, Y. *et al.* (2008b) Evaluating the ROC performance of markers for future events. *Lifetime Data Analysis*, **14**, 86–113.

Peri, S., Navarro, J.D., Amanchy, R. *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome Research*, **13**, 2363–2371.

Perry, G.H., Ben-Dor, A., Tsalenko, A. *et al.* (2008) The fine-scale and complex architecture of human copy-number variation. *American Journal of Human Genetics*, **82**, 685–695.

Pettersson, F., Morris, A.P., Barnes, M.R. and Cardon, L.R. (2008) Goldsurfer2 (Gs2): a comprehensive tool for the analysis and visualization of genome wide association studies. *BMC Bioinformatics*, **9**, 138.

Pirooznia, M., Yang, J.Y., Yang, M.Q. and Deng, Y. (2008) A comparative study of different machine learning methods on microarray gene expression data. *BMC Genomics*, **9** (Suppl 1), S13.

Pisitkun, T., Shen, R.F. and Knepper, M.A. (2004) Identification and proteomic profiling of exosomes in human urine. *Proceedings of the National Academy of Sciences of the United States of America*, **101**, 13368–13373.

Popper, S., Shimizu, C., Shike, H. *et al.* (2007) Gene-expression patterns reveal underlying biological processes in Kawasaki disease. *Genome Biology*, **8**, R261.

Prohl, J., Röther, J., Kluge, S. *et al.* (2007) Prediction of short-term and long-term outcomes after cardiac arrest: a prospective multivariate approach combining biochemical, clinical, electrophysiological, and neuropsychological investigations. *Critical Care Medicine*, **35** (5), 1230–1237.

Ptitsyn, A.A., Weil, M.M. and Thamm, D.H. (2008) Systems biology approach to identification of biomarkers for metastatic progression in cancer. *BMC Bioinformatics*, **9** (Suppl 9), S8.

Pujana, M.A., Han, J.D., Starita, L.M. *et al.* (2007) Network modeling links breast cancer susceptibility and centrosome dysfunction. *Nature Genetics*, **39**, 1338–1349.

Purcell, S., Neale, B., Todd-Brown, K. *et al.* (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. *American Journal of Human Genetics*, **81**, 559–575.

Purohit, P.V. and Rocke, D.M. (2003) Discriminant models for highthroughput proteomics mass spectrometer data. *Proteomics*, **3**, 1699–1703.

Quackenbush, J. (2006) Microarray analysis and tumor classification. *The New England Journal of Medicine*, **354**, 2463–2472.

Quinlan, J.R. (1993) *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, USA.

Radulovic, D., Jelveh, S., Ryu, S. *et al.* (2004) Informatics platform for global proteomic profiling and biomarker discovery using liquid chromatography-tandem mass spectrometry. *Molecular & Cellular Proteomics*, **3**, 984–997.

Rajan, S., Williams, S.S., Jagatheesan, G. *et al.* (2006) Microarray analysis of gene expression during early stages of mild and severe cardiac hypertrophy. *Physiological Genomics*, **27**, 309–317.

Ransohoff, D.F. (2005) Bias as a threat to the validity of cancer molecular-marker research. *Nature Reviews. Cancer*, **5**, 142–149.

Rao, S.R. and Schoenfeld, D.A. (2007) Survival methods. *Circulation*, **115**, 109–113.

Rapaport, F., Barillot, E. and Vert, J.P. (2008) Classification of arrayCGH data using fused SVM. *Bioinformatics (Oxford, England)*, **24**, i375–i382.

Raza, S., Robertson, K.A., Lacaze, P.A. *et al.* (2008) A logic-based diagram of signalling pathways central to macrophage activation. *BMC Systems Biology*, **2**, 36.

Redon, R., Ishikawa, S., Fitch, K.R. *et al.* (2006) Global variation in copy number in the human genome. *Nature*, **444**, 444–454.

Ressom, H.W., Varghese, R.S., Zhang, Z. *et al.* (2008) Classification algorithms for phenotype prediction in genomics and proteomics. *Frontiers in Bioscience: A Journal and Virtual Library*, **13**, 691–708.

Rhodes, D.R., Kalyana-Sundaram, S., Mahavisno, V. *et al.* (2007) Oncomine 3.0: genes, pathways, and networks in a collection of 18,000 cancer gene expression profiles. *Neoplasia (New York, NY)*, **9**, 166–180.

Ringnér, M. (2008) What is principal component analysis? *Nature Biotechnology*, **26**, 303–304.

Rives, A.W. and Galitski, T. (2003) Modular organization of cellular networks. *Proceedings of the National Academy of Sciences of the United States of America*, **100**, 1128–1133.

Russell, S. and Norvig, P. (2002) *Artificial Intelligence: A Modern Approach*, 2nd edn, Prentice Hall.

Ruttenberg, A., Clark, T., Bug, W. *et al.* (2007) Advancing translational research with the Semantic Web. *BMC Bioinformatics*, **8** (Suppl 3), S2.

Sabatine, M.S., Liu, E., Morrow, D.A. *et al.* (2005) Metabolomic identification of novel biomarkers of myocardial ischemia. *Circulation*, **112**, 3868–3875.

Saeys, Y., Inza, I. and Larrañaga, P. (2007) A review of feature selection techniques in bioinformatics. *Bioinformatics (Oxford, England)*, **23** (19), 2507–2517.

Sanoudou, D., Vafiadaki, E., Arvanitis, D.A. *et al.* (2005) Array lessons from the heart: focus on the genome and transcriptome of ardiomyopathies. *Physiological Genomics*, **14**, 131–143.

Sawyers, C.L. (2008) The cancer biomarker problem. *Nature*, **452**, 548–552.

Schachtner, R., Lutter, D., Knollmüller, P. *et al.* (2008) Knowledge-based gene expression classification via matrix factorization. *Bioinformatics (Oxford, England)*, **24**, 1688–1697.

Schäfer, J. and Strimmer, K. (2005) A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statistical Applications in Genetics and Molecular Biology*, **4**, Article32.

Scholz, M. and Fiehn, O. (2007) SetupX: A public study design database for metabolomic projects. *Pacific Symposium on Biocomputing*, **12**, 169–180.

Seo, D., Ginsburg, G.S. and Goldschmidt-Clermont, P.J. (2006) Gene expression analysis of cardiovascular diseases: novel insights into biology and clinical applications. *Journal of the American College of Cardiology*, **48**, 227–235.

Shannon, P., Markiel, A., Ozier, O. *et al.* (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498.

Shapiro, D.E. (1999) The interpretation of diagnostic tests. *Statistical Methods in Medical Research*, **8**, 113–134.

Sharma, U.C., Pokharel, S., Evelo, C.T. and Maessen, J.G. (2005) A systematic review of large scale and heterogeneous gene array data in heart failure. *Journal of Molecular and Cellular Cardiology*, **38**, 425–432.

Shulaev, V. (2006) Metabolomics technology and bioinformatics. *Briefings in Bioinformatics*, **7**, 128–139.

Simon, R. (2006) Development and evaluation of therapeutically relevant predictive classifiers using gene expression profiling. *Journal of the National Cancer Institute*, **98**, 1169–1171.

Simon, R. (2008) The use of genomics in clinical trial design. *Clinical Cancer Research*, **14**, 5984–5993.

Skalak, D. (1994) Prototype and feature selection by sampling and random mutation hill-climbing algorithms. Proceedings of the Eleventh International Conference on Machine Learning, New Brunswick, New Jersey, pp. 293–301.

Smidt, N., Rutjes, A.W., van der Windt, D.A. *et al.* (2006b) The quality of diagnostic accuracy studies since the STARD statement: has it improved? *Neurology*, **67**, 792–797.

Smidt, N., Rutjes, A.W., van der Windt, D.A. *et al.* (2006a) Reproducibility of the STARD checklist: an instrument to assess the quality of reporting of diagnostic accuracy studies. *BMC Medical Research Methodology*, **6**, 12.

Smith, A., Cheung, K., Krauthammer, M. *et al.* (2007) Leveraging the structure of the Semantic Web to enhance information retrieval for proteomics. *Bioinformatics (Oxford, England)*, **23** (22), 3073–3079.

Smith, C.A., O'Maille, G., Want, E.J. *et al.* (2006) METLIN: a metabolite mass spectral database. *Therapeutic Drug Monitoring*, **27**, 747–751.

Smith, I.C. and Baert, R. (2003) Medical diagnosis by high resolution NMR of human specimens. *IUBMB Life*, **55**, 273–277.

Smulders, Y.M., Thijs, A. and Twisk, J.W. (2008) New cardiovascular risk determinants do exist and are clinically useful. *European Heart Journal*, **29**, 436–440.

Sodeck, G.H., Domanovits, H., Sterz, F. *et al.* (2007) Can brain natriuretic peptide predict outcome after cardiac arrest? An observational study. *Resuscitation*, **74** (3), 439–445.

Sörnmo, L. and Laguna, P. (2005) *Bioelectrical Signal Processing in Cardiac and Neurological Applications*, Academic Press, Burlington, MA, USA.

Sotomayor, B. and Childers, L. (2005) *Globus® Toolkit 4: Programming Java Services*, Morgan Kaufmann, San Francisco, USA.

Spinale, F.G. (2004) Matrix metalloproteinase gene polymorphisms in heart failure: new pieces to the myocardial matrix puzzle. *European Heart Journal*, **25**, 631–633.

SPSS Inc. (2008) [www.spss.com], last accessed 21 May 2008.

Sreekumar, A., Poisson, L.M., Rajendiran, T.M. *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, **457**, 910–914.

Stammet, P., Beyenburg, S., Codreanu, A. *et al.* (2008) "Biosignature to predict outcome after hypothermia for patients surviving cardiac arrest (NORTH POLE STUDY)".

Stanton, L.W., Garrard, L.J., Damm, D. *et al.* (2000) Altered patterns of gene expression in response to myocardial infarction. *Circulation Research*, **86**, 939–945.

STARD Statement (2009) [http://www.stard-statement.org], last accessed: 9 April 2009.

Statnikov, A., Aliferis, C.F., Tsamardinos, I. *et al.* (2005) A comprehensive evaluation of multicategory classification methods for microarray gene expression cancer diagnosis. *Bioinformatics (Oxford, England)*, **21**, 631–643.

Statnikov, A., Wang, L. and Aliferis, C.F. (2008) A comprehensive comparison of random forests and support vector machines for microarray-based cancer classification. *BMC Bioinformatics*, **9**, 319.

Stein, L.D. (2008) Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges. *Nature Reviews. Genetics*, **9**, 678–688.

Stenson, P.D., Ball, E.V., Mort, M. *et al.* (2003) Human gene mutation database (HGMD). *Human Mutation*, **21**, 577–581.

Sterne, J.A. and Davey Smith, G. (2001) Sifting the evidence-what's wrong with significance tests? *BMJ (Clinical Research Ed)*, **322**, 226–231.

Steuer, R., Kurths, J., Daub, C.O. *et al.* (2002) The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics (Oxford, England)*, **18** (Suppl 2), S231–S240.

Strobl, C., Boulesteix, A.L., Zeileis, A. and Hothorn, T. (2007) Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics*, **8**, 25.

Stroup, D.F., Berlin, J.A., Morton, S.C. *et al.* (2000) Meta-analysis of observational studies in epidemiology: a proposal for reporting. Meta-analysis of observational studies in epidemiology (MOOSE) group. *The Journal of the American Medical Association*, **283**, 2008–2012.

Subramanian, A., Tamayo, P., Mootha, V.K. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 15545–15550.

Sullivan, L.M. (2006) Estimation from samples. *Circulation*, **114** (5), 445–449.

Suthram, S., Sittler, T. and Ideker, T. (2005) The Plasmodium protein network diverges from those of other eukaryotes. *Nature*, **438**, 108–112.

Swets, J.A. (1988) Measuring the accuracy of diagnostic systems. *Science*, **240**, 1285–1293.

Sysi-Aho, M., Katajamaa, M., Yetukuri, L. and Oresic, M. (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. *BMC Bioinformatics*, **8**, 93.

Tan, F.L., Moravec, C.S., Li, J. *et al.* (2002) The gene expression fingerprint of human heart failure. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 11387–11392.

Tang, J., Tan, C.Y., Oresic, M. and Vidal-Puig, A. (2009) Integrating post-genomic approaches as a strategy to advance our understanding of health and disease. *Genome Medicine*, **1**, 35.

Tang, T., Pankow, J.S., Carr, J.J. *et al.* (2007) Association of sICAM-1 and MCP-1 with coronary artery calcification in families enriched for coronary heart disease or hypertension: the NHLBI Family Heart Study. *BMC Cardiovascular Disorders*, **7**, 30.

Tang, Y., Nee, A.C., Lu, A. *et al.* (2003) Blood genomic expression profile for neuronal injury. *Journal of Cerebral Blood Flow and Metabolism*, **23**, 310–319.

Tanriverdi, K. and Freedman, J.E. (2008) Blood and cardiovascular disease: The promise and limitations of gene expression analysis. *Circulation: Cardiovascular Genetics*, **1**, 7–9.

Taraban, R. and Blanton, R.L. (eds) (2008) *Creating Effective Undergraduate Research Programs in Science: The Transformation from Student to Scientist*, Teachers College Press, New York, USA.

Taylor, I.W., Linding, R., Warde-Farley, D. *et al.* (2009) Dynamic modularity in protein interaction networks predicts breast cancer outcome. *Nature Biotechnology*, **27**, 199–204.

The CONSORT Statement (2009) [www.consort-statement.org], last accessed: 9 April 2009.

The Equator Network (2009) [www.equator-network.org], last accessed: 9 April 2009.

The hypothermia after cardiac arrest study group (2000) Mild therapeutic hypothermia to improve the neurologic outcome after cardiac arrest. *The New England Journal of Medicine*, **346**, 549–556.

The International HapMap Consortium (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

The International HapMap Consortium (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.

Thompson, A., Brennan, K., Cox, A. *et al.* (2008) Evaluation of the current knowledge limitations in breast cancer research: a gap analysis. *Breast Cancer Research*, **10**, R26.

Tian, L., Greenberg, S.A., Kong, S.W. *et al.* (2005) Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 13544–13549.

Tibshirani, R., Hastie, T., Narasimhan, B. and Chu, G. (2002) Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences of the United States of America*, **99**, 6567–6572.

Tibshirani, R., Hastie, T., Narasimhan, B. *et al.* (2004) Sample classification from protein mass spectrometry, by 'peak probability contrasts'. *Bioinformatics (Oxford, England)*, **20**, 3034–3044.

Timofeeva, A.V., Goryunova, L.E., Khaspekov, G.L. *et al.* (2006) Altered gene expression pattern in peripheral blood leukocytes from patients with arterial hypertension. *Annals of the New York Academy of Sciences*, **1091**, 319–335.

Topol, E.J., Smith, J., Plow, E.F. and Wang, Q.K. (2006) Genetic susceptibility to myocardial infarction and coronary artery disease. *Human Molecular Genetics*, **15**, Spec No 2: R117–R123.

Tranchevent, L.C., Barriot, R., Yu, S. *et al.* (2008) ENDEAVOUR update: a web resource for gene prioritization in multiple species. *Nucleic Acids Research*, **36** (Web Server issue), W377–W384.

Tusher, V.G., Tibshirani, R. and Chu, G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences of the United States of America*, **98**, 5116–5121.

van de Vijver, M.J., He, Y.D., van't Veer, L.J. *et al.* (2002) A gene-expression signature as a predictor of survival in breast cancer. *The New England Journal of Medicine*, **347**, 1999–2009.

Van Dien, S. and Schilling, C.H. (2006) Bringing metabolomics data into the forefront of systems biology. *Molecular Systems Biology*, **2**, 2006.0035.

Van, Q.N. and Veenstra, T.D. (2009) How close is the bench to the bedside? Metabolic profiling in cancer research. *Genome Medicine*, **1**, 5.

van't Veer, L.J., Dai, H., van de Vijver, M.J. *et al.* (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.

van't Veer, L.J. and Bernards, R. (2008) Enabling personalized cancer medicine through analysis of gene-expression patterns. *Nature*, **452**, 564–570.

Varma, S. and Simon, R. (2006) Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, **7**, 91.

Vasan, R.S. (2006) Biomarkers of cardiovascular disease: molecular basis and practical considerations. *Circulation*, **113** (19), 2335–2362.

Veltri, P. (2008) Algorithms and tools for analysis and management of mass spectrometry data. *Briefings in Bioinformatics*, **9**, 144–155.

Vergara, I.A., Norambuena, T., Ferrada, E. *et al.* (2008) StAR: a simple tool for the statistical comparison of ROC curves. *BMC Bioinformatics*, **9**, 265.

Vicens, Q. and Bourne, P.E. (2007) Ten simple rules for a successful collaboration. *PLoS Computational Biology*, **3**, e44.

Vodovotz, Y., Csete, M., Bartels, J. *et al.* (2008) Translational systems biology of inflammation. *PLoS Computational Biology*, **4**, e1000014.

Wachi, S., Yoneda, K. and Wu, R. (2005) Interactome-transcriptome analysis reveals the high centrality of genes differentially expressed in lung cancer tissues. *Bioinformatics (Oxford, England)*, **21**, 4205–4208.

Wagner, A. (2005) *Robustness and Evolvability in Living Systems*, Princeton University Press, Princeton, New Jersey, USA.

Wang, T.J. (2008) New cardiovascular risk factors exist, but are they clinically useful? *European Heart Journal*, **29**, 441–444.

Wang, H., Zheng, H. and Azuaje, F. (2008) Clustering-based approaches to SAGE data mining. *BioData Mining*, **1**, 5.

Wang, H., Zheng, H., Simpson, D. and Azuaje, F. (2006) Machine learning approaches to supporting the identification of photoreceptor-enriched genes based on expression data. *BMC Bioinformatics*, **7**, 116.

Wang, Y., Barbacioru, C.C., Shiffman, D. *et al.* (2007) Gene expression signature in peripheral blood detects thoracic aortic aneurysm. *PLoS ONE*, **2** (10), e1050.

Wang, Y., Klijn, J.G., Zhang, Y. *et al.* (2005) Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer. *Lancet*, **365**, 671–679.

Watkinson, J., Wang, X., Zheng, T. and Anastassiou, D. (2008) Identification of gene interactions associated with disease from gene expression data using synergy networks. *BMC Systems Biology*, **2**, 10.

Webb-Robertson, B.J. and Cannon, W.R. (2007) Current trends in computational inference from mass spectrometry-based proteomics. *Briefings in Bioinformatics*, **8**, 304–317.

Westfall, P.H., Young, S.S. and Wright, S.P. (1993) On adjusting P-values for multiplicity. *Biometrics*, **49**, 941–945.

Whiting, P., Rutjes, A.W., Reitsma, J.B. *et al.* (2003) The development of QUADAS: a tool for the quality assessment of studies of diagnostic accuracy included in systematic reviews. *BMC Medical Research Methodology*, **3**, 25.

Whiting, P.F., Sterne, J.A., Westwood, M.E. *et al.* (2008) Graphical presentation of diagnostic information. *BMC Medical Research Methodology*, **8**, 20.

Wilkinson, M.D. and Links, M. (2002) BioMOBY: an open source biological web services proposal. *Briefings in Bioinformatics*, **3**, 331–341.

Willer, C.J., Sanna, S., Jackson, A.U. *et al.* (2008) Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genetics*, **40**, 161–169.

Willinsky, J. (2006) *The Access Principle: The Case for Open Access to Research and Scholarship*, MIT Press, Cambridge, MA, USA.

Wilson, I.D., Plumb, R., Granger, J. *et al.* (2005) HPLC-MS-based methods for the study of metabonomics. *Journal of Chromatography. B, Analytical Technologies in the Biomedical and Life Sciences*, **817**, 67–76.

Winter, A., Funkat, G., Haeber, A. *et al.* (2007) Integrated information systems for translational medicine. *Methods of Information in Medicine*, **46**, 601–607.

Wishart, D.S. (2007) Current progress in computational metabolomics. *Briefings in Bioinformatics*, **8**, 279–293.

Wishart, D.S. and Greiner, R. (2007) Computational approaches to metabolomics: an introduction. *Pacific Symposium on Biocomputing*, **12**, 112–114.

Wishart, D.S., Tzur, D., Knox, C. *et al.* (2007) HMDB: the human metabolome database. *Nucleic Acids Research*, **35**, D521–D526.

Wittchen, F., Suckau, L., Witt, H. *et al.* (2006) Genomic expression profiling of human inflammatory cardiomyopathy (DCMi) suggests novel therapeutic targets. *Journal of Molecular Medicine (Berlin, Germany)*, **85**, 257–271.

Witten, I.H. and Frank, E. (2005) *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd edn, Morgan Kaufmann, San Francisco, USA.

Wittke-Thompson, J.K., Pluzhnikov, A. and Cox, N.J. (2005) Rational inferences about departures from Hardy-Weinberg equilibrium. *American Journal of Human Genetics*, **76**, 967–986.

Wood, I.A., Visscher, P.M. and Mengersen, K.L., (2007) Classification based upon gene expression data: bias and precision of error rates. *Bioinformatics (Oxford, England)*, **23**, 1363–1370.

Xia, J., Bjorndahl, T.C., Tang, P. and Wishart, D.S. (2008) MetaboMiner–semi-automated identification of metabolites from 2D NMR spectra of complex biofluids. *BMC Bioinformatics*, **9**, 507.

Xie, Q., Ratnasinghe, L.D., Hong, H. *et al.* (2005) Decision forest analysis of 61 single nucleotide polymorphisms in a case-control study of esophageal cancer. *BMC Bioinformatics*, **6** (Suppl 2), S4.

Xu, M., Kao, M.C., Nunez-Iglesias, J. *et al.* (2008) An integrative approach to characterize disease-specific pathways and their coordination: a case study in cancer. *BMC Genomics*, **9** (Suppl 1), S12.

Xu, X., Veenstra, T.D., Fox, S.D. *et al.* (2005) Measuring fifteen endogenous estrogens simultaneously in human urine by high-performance liquid chromatography-mass spectrometry. *Analytical Chemistry*, **77**, 6646–6654.

Yang, Y., Pospisil, P., Iyer, L.K. *et al.* (2008) Integrative genomic data mining for discovery of potential blood-borne biomarkers for early diagnosis of cancer. *PLoS ONE*, **3** (11), e3661.

Yusuf, S., Hawken, S., Ounpuu, S. *et al.* (2004) Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case–control study. *The Lancet*, **364**, 937–952.

Zethelius, B., Berglund, L., Sundström, J. *et al.* (2008) Use of multiple biomarkers to improve the prediction of death from cardiovascular causes. *The New England Journal of Medicine*, **358**, 2107–2116.

Zhang, B. and Horvath, S. (2005) A general framework for weighted gene co-expression network analysis. *Statistical Applications in Genetics and Molecular Biology*, **4**, Article17.

Zhang, Y., Zhang, Y., Adachi, J. *et al.* (2007a) MAPU: Max-Planck Unified database of organellar, cellular, tissue and body fluid proteomes. *Nucleic Acids Research*, **35** (Database issue), D771–D779.

Zhang, Z., Chen, D. and Fenstermacher, D.A. (2007b) Integrated analysis of independent gene expression microarray datasets improves the predictability of breast cancer outcome. *BMC Genomics*, **8**, 331.

Zhu, J., Sanborn, J.Z., Benz, S. *et al.* (2009) The UCSC Cancer Genomics Browser. *Nature Methods*, **6**, 239–240.

Ziegler, A., König, I.R. and Thompson, J.R. (2008) Biostatistical aspects of genome-wide association studies. *Biometrical Journal*, **50**, 8–28.

Zou, K.H., O'Malley, A.J. and Mauri, L. (2007) Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models. *Circulation*, **115**, 654–657.

# Index

*This index was prepared by Neil Manley.*