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’

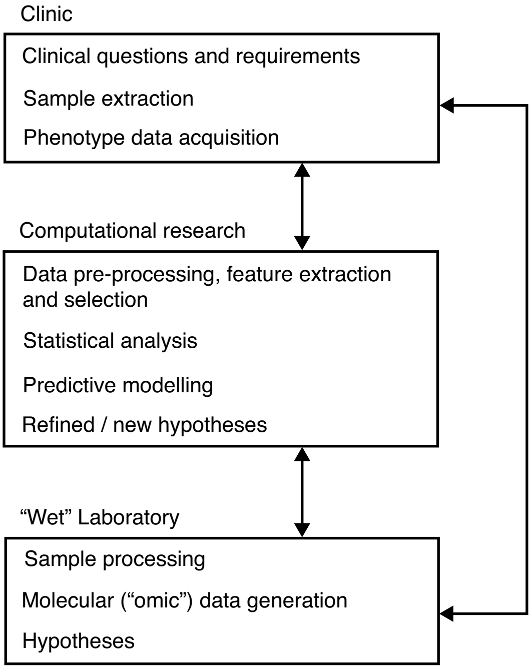

Figure 1.1 The dialogue between clinical, laboratory and computational research environments in the context of translational biomedical research. Examples of typical tasks and applications

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

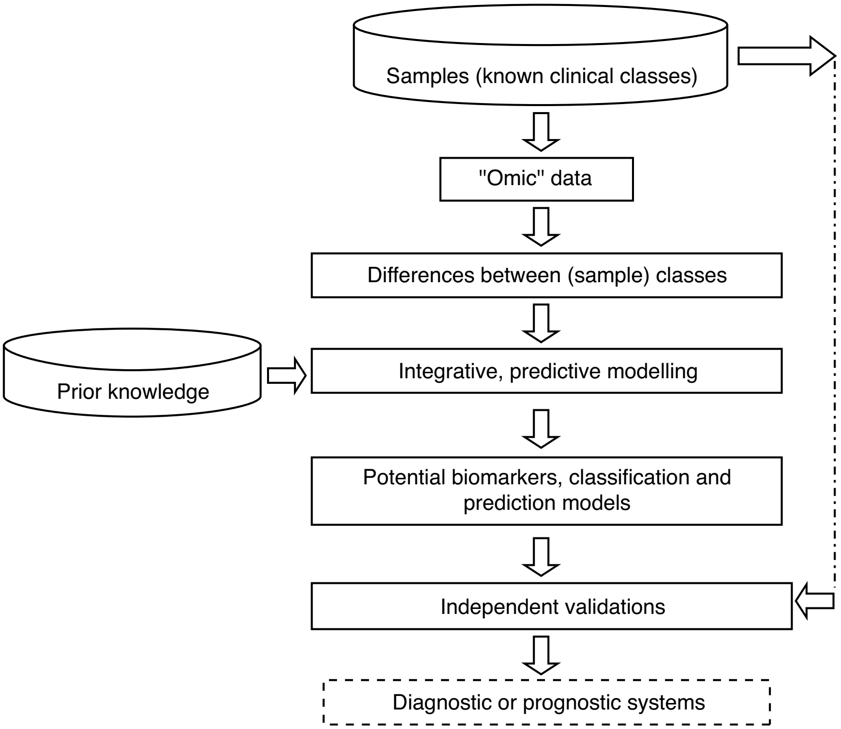
Samples (known clinical classes)
"Omic" data
Differences between (sample) classes
Prior knowledge
Integrative, predictive modelling
Potential biomarkers, classification and prediction models
Independent validations
Diagnostic or prognostic systems

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.

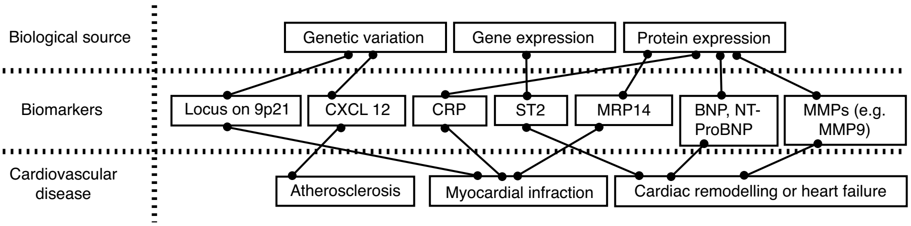

Figure 1.3 Examples of cardiovascular biomarkers and their relationship to different 'omic' technologies and diseases (Vasan, 2006; Gerszten and Wang, 2008)

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 pathophysiology 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.