PREFACE

By combining data mining, statistics, and computer science with computational biology, bioinformatics, genomics, proteomics and personalized medicine, and then adding other familiar and emerging fields we will portray an inclusive interdisciplinary environment in which a growing number of us enjoy working. This book is written for students, practitioners, and researchers involved in biomedical or data mining projects that use gene expression or protein expression data. This group includes students of bioinformatics, data mining, computational biology, biomedical sciences, and other related areas. In addition, this book is directed to anyone interested in efficient methods of knowledge discovery based on data sets consisting of thousands or millions of variables and often only dozens or hundreds of observations, whether the data is related to biomedical research or to such other fields as market, financial, or physical modeling.

Even experienced data miners may be surprised to learn that some well established methods and procedures fail in such high-dimensional $p \gg N$ situations. New methods are constantly being developed, making it even more difficult for biomedical researchers and bioinformaticians to select appropriate methods to use for a particular project. No single method works optimally in all situations. Some approaches are efficacious for some study goals but inappropriate for others. For example, we should not drive biomarker discovery with unsupervised learning methods. However, methods like clustering are so popular that they are used indiscriminately, not only for studies seeking new taxonomic information, for which they are a perfect choice, but also often as primary methods in situations in which supervised approaches should be used. The confusion among biomedical researchers is visible in the large body of papers describing experiments based on methods that are inappropriate for particular study goals or for the data at hand. Such experiments provide either misleading conclusions or only anecdotal evidence.

One of the primary goals of this book is to provide clear guidance on when to use which methods and why. Such subjects as why some of the popular approaches can lead to inferior solutions, and sometimes even to the worst possible results will be examined. Among the topics discussed are: univariate versus multivariate approaches, supervised versus unsupervised methods, selecting proper methods for feature selection, regularization of biomarker discovery leading to more stable and generalizable classifiers, ensemble-based estimation of the misclassification error rate, identification of the *Informative Set of Genes* containing all information significant for class differentiation. The book covers all aspects of gene and protein expression analysis—technology, data preprocessing and quality assessment, basic exploratory analysis, unsupervised and supervised learning algorithms. Special attention is

given to multivariate biomarker discovery leading to parsimonious and generalizable classifiers.

Students of my course on *Data Mining for Genomics and Proteomics*, a part of the Central Connecticut State University (CCSU) data mining program, have diverse backgrounds, from molecular biology PhDs to those with no biology background. This book is written for an equally broad audience. The only assumption is that the reader is interested in the subject. As a result, for different readers different parts of this book may be seen as introductory and different parts as advanced. All efforts have been made to start each subject from its basics.

The book comprises five main chapters with the sixth chapter containing examples of hands-on analysis of real-world data sets. Chapter 1 provides an introduction to basic, mostly biological, terms. Chapter 2 focuses on microarray technology and basic analysis of gene expression data, including low-level preprocessing methods, probe set level preprocessing and filtering, basic exploratory analysis, and such unsupervised learning methods as cluster analysis, principal component analysis, and self-organizing maps.

Chapters 3 and 4 are the core of the book, covering multivariate methods for feature selection, biomarker discovery, and identification of generalizable and interpretable classifiers. Such biomarkers and classifiers can be designed and used for early medical diagnosis, for tailoring therapy selection to prediction of individual response to available treatment modalities, for assessing treatment progression, for drug discovery, and in any other areas that can benefit from parsimonious multivariate markers identified from a very large number of variables.

Chapter 3 starts with an overview of biomarker discovery and classification, followed by the coverage of feature selection methods, and then descriptions of selected supervised learning algorithms. Important differences between univariate and multivariate approaches as well as between supervised and unsupervised methods are discussed. Three learning algorithms are described in detail. *Linear discriminant analysis* represents classical and still powerful parametric methods that should be in the portfolio of any data miner and bioinformatician. *Support vector machines* are newer but already well-established methods for designing both linear and nonlinear classifiers. *Random forests* represent recent ensemble-based approaches. In addition, two other learning algorithms are described—*k-nearest neighbors* and *artificial neural networks*, useful in some situations, even if they are usually not the first choice in biomarker discovery. A discussion of the bootstrap and ensemble-based approaches culminates with defining the *modified bagging* schema that allows for building ensembles of classifiers based on randomized training sets generated by stratified sampling without replacement.

Chapter 4 defines the *Informative Set of Genes* as a set containing all information significant for the differentiation of classes represented in training data, which can be used in two complementary capacities. First, it facilitates biological interpretation of class differences. Second, in combination with ensembles of classifiers generated with the use of the *modified bagging* schema, it allows the identification of robust biomarkers. The method introduced in Chapter 4 allows for the optimization of feature selection that leads to identification of parsimonious and generalizable multivariate biomarkers with plausible biological interpretation.

Chapter 5 addresses the analysis of protein expression data. Data mining methods for feature selection, biomarker discovery and classification are the same in proteomics as those described, in Chapters 3 and 4, for genomics. For that reason, Chapter 5 focuses on proteomic technologies and on the analytical steps that are intrinsic to proteomics. The covered technologies include protein microarrays, two-dimensional gel electrophoresis and MALDI-TOF and SELDI-TOF mass spectrometry. Among the discussed analytical methods are: preprocessing of mass spectrometry data, and associating biomarker peaks with proteins.

The exercises included in Chapters 2–4 give readers an opportunity to put the methods they have learned to practical use. To make it easier and more enjoyable, many of these exercises are covered by studies presented in Chapter 6.

Finally, Chapter 6 provides examples of the analysis of real-world gene expression data sets that illustrate the practical application of the methods described in this book. The examples illustrate such tasks as designing a multi-stage classification schema, combining two gene expression data sets into one training set, identification of the *Informative Set of Genes* and its primary expression patterns, and using ensembles of classifiers built with the *modified bagging* schema to identify parsimonious and generalizable biomarkers.

DARIUS M. DZIUDA

Bethany, Connecticut
June 2009