Preface

In recent times, there has been an explosion in the development of comprehensive, high-throughput methods for molecular biology experimentation. An example is the advent in DNA microarray technologies, such as cDNA arrays and oligonucleotide arrays, that provide means for measuring tens of thousands of genes simultaneously. These technologies benefit biological research greatly and further our understanding of biological processes by drawing together researchers in biology and quantitative fields including statistics, mathematics, computer science, and physics. In addition to the enormous scientific potential of microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research.

This book has been written with two types of readers in mind: biologists who will undertake the statistical analyses of their own experimental microarray data, and biostatisticians entering the field of microarray gene expression data analysis. The primary focus of the book is on data analysis methods for this field; however, the biology and technology behind gene expression microarray experiments, as well as cleaning and normalization of the data, will be briefly covered.

Although biological experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels, organized by genes versus tissue samples. In the case where a tissue sample corresponds to a single microarray experiments, we can represent the output from M experiments in the form of a $N \times M$ array (matrix). Each column of the matrix (the expression signature vector) contains the expression levels on the N genes monitored in the microarrays, while each row (the expression profile) contains the expression levels of a gene as it varies over the M tissue samples. Outside this matrix of expression levels, we may have covariate information for samples, genes, or both. The goal of microarray data analysis is to make inferences among samples, genes, and their expression levels and covariates.

The actual measurement of the expression levels raises several statistical issues in experimental design, image processing, outlier detection, transformations, and nonlinear modeling. We consider some of these issues (which are still ongoing as we complete this book) in the first two chapters. The rest of the book then considers the analysis of the microarray data, assuming that they have been appropriately preprocessed.

This analysis is centered on methods for the detection of differential expression, for cluster analysis (unsupervised classification), and for discriminant analysis (supervised classification) of microarray data.

An important and common question in microarray experiments is the detection of genes that are differentially expressed in tissue samples across a number of specified classes. These classes may correspond to tissues (cells) that are at different stages in some process, in distinct pathological states, or under different experimental conditions. A plethora of methods to detect differential gene expression are presented.

Cluster analyses have demonstrated their utility in the elucidation of unknown gene function, the validation of gene discoveries, and the interpretation of biological processes. Discriminant analysis is playing an ever-increasing role in predicting gene function classes and cancer classification.

There are two distinct clustering problems with microarray data. One problem concerns the clustering of the tissues on the basis of the genes. The clusters of tissues can play a useful role in the discovery and understanding of new subclasses of diseases. The second problem concerns the clustering of the genes on the basis of the tissues. The clusters of genes obtained can be used to search for genetic pathways or groups of genes that might be regulated together. Also, in the first problem above, we may wish first to summarize the information in the very large number of genes by clustering them into groups, which can be represented by some metagenes. We can then carry out the clustering of the tissues in terms of these metagenes.

In both the clustering of the tissues and the genes, hierarchical (agglomerative) clustering has been the most widely used method for the analysis of patterns of gene expression. It produces a representation of the data with the shape of a binary tree, in which the most similar patterns are clustered in a hierarchy of nested subsets. Nevertheless, classical hierarchical clustering presents drawbacks when dealing with data containing a non-negligible amount of noise, as is the present case. Also, there is no reason why the clusters of tissues or genes should belong to a hierarchy such as in the evolution of species. In this book, the emphasis is on a model-based approach to clustering. An advantage of model-based clustering is that it provides a sound mathematical framework for clustering. In particular, it provides a principled statistical approach to the practical questions that arise in applying clustering methods, namely, the question of what metric (distance function) to adopt and the question of how many clusters there are in the data.

In recent times, model-based clustering has become very popular in the statistical literature. Unfortunately, as the data to be analyzed from microarray experiments often have gene-to-sample ratios of approximately 100-fold, off-the-shelf parametric methodology does not apply at least to the classification of the tissues on the basis of the genes. This is because the dimension of the feature space (the number of genes)

is so much greater than the number of observations (the number of tissues). But even the cluster analysis of the genes on the basis of the tissues is a nonstandard problem, as the genes are not all independently distributed.

An obvious way to handle the very large number of genes is to perform a principal component analysis (PCA) and carry out the cluster analysis on the basis of the leading components. But a potential problem with a PCA is the determination of an appropriate number of principal components (PCs) useful for clustering. A common practice is to choose the first few leading components. But it is not clear where to stop and whether some of these components are caused by some artifact or noises in the data unrelated to the clustering task. Also, there is the difficulty of interpretation of components because each component has loadings generally on all genes.

Hence the focus in the book is on the EMMIX-GENE procedure, which is a normal mixture-based method of clustering that has been especially developed for the clustering of tissue samples or other high-dimensional data. This procedure has an option for an initial selection of the genes where genes that appear to have little clustering capacity are discarded. It then clusters the (standardized) gene profiles into groups, effectively using Euclidean distance as the metric, with the aim that highly correlated genes are put in the same cluster. Each group of genes is then represented by a single metagene (the group-sample mean) and then clustering is performed in terms of the metagenes. This divide-and-conquer approach is becoming popular in the bioinformatics literature for both unsupervised and supervised classification of tissue samples.

The clustering step of EMMIX-GENE makes use of, if needed, mixtures of factor analyzers. That is, it provides a global nonlinear approach to dimension reduction as it postulates a finite mixture of linear submodels (factor models) for the distribution of the full signature vector or a reduced version of metagenes given the (unobservable) factors. Thus, it is a local dimensionality reduction method in contrast with a PCA, which is a global linear method.

A number of discriminant rules are discussed for the supervised classification of the tissue samples. However, the aim of this book is not to provide a comprehensive review of available methods but rather to focus on what we think are useful methods for the analysis of microarray data. To this end, the focus in discriminant analysis of tissue samples is on the support vector machine. It has the advantage that it can be formed from all the genes and its performance is generally not too disadvantaged as a consequence of using all the genes. Its performance can be improved by undertaking feature selection using an easily implemented procedure called recursive feature elimination.

In the statistical analyses, including discriminant and cluster analyses, some form of feature selection will usually be carried out. A consequence of basing the final analysis on a selected "top" subset of the available genes is that there will typically be a selection bias that needs to be corrected for in relating the conclusions to subsequent (new) data. In the case of a discriminant rule, it means that the selection bias has to be allowed for in the estimation of the generalization error. Otherwise, a false overoptimistic impression will be obtained for the discriminatory power of the rule.

This bias has often been overlooked in the bioinformatics literature. Also, this bias arises in an unsupervised context with tests and plots on the number of clusters.

The first two chapters of this book aim to (1) provide a bridge to the biological and technical aspects involved in microarray experiments, and (2) summarize and emphasize the need for basic research in DNA array technologies and statistical thinking through every step of the microarray experiment and analysis to enhance reliability and reproducibility of research results.

Chapter 1 is an introductory chapter and provides a review on DNA microarrays and relevant technology. In particular, we begin with the biological principles behind microarray experiments. Background information on the substrates and technology used in microarray gene expression studies is intended for the biostatistician who is not familiar with the biological experiments. We discuss DNA, cDNA, oligonucleotides, and the development of microarray technology, as well as the steps involved in the manufacture of cDNA microarrays and in generating experimental microarray data. Commercial arrays, primarily the GeneChip®, are also briefly introduced in this chapter.

Chapter 2 discusses cleaning and normalization of gene expression microarray data, as well as the need for designs of experiments with replicated data.

Chapter 3 considers in a general context some methods for the cluster analysis of multivariate data consisting of n independent observations taken on a p-dimensional feature vector associated with the random phenomenon of interest. The focus is on model-based methods of clustering and it covers the use of mixtures of factor analyzers for high-dimensional data such as microarray data. In relation to the problem of how many clusters there are in the data, consideration is given to the problem of assessing the number of components in a mixture model by resampling.

Chapter 4 considers the development of the model-based methodology covered in Chapter 3 for its application to problems in the clustering of tissue samples. The emphasis is on the EMMIX-GENE procedure which has been developed specifically for the clustering of tissue samples. Its application to real microarray data sets is illustrated on two well-known sets in the literature. Also, it is demonstrated on several real data sets how this model-based approach to clustering can be used to consider the question of how many clusters of tissues there are in the data.

Chapter 5 focuses on the selection of differentially expressed genes in known classes of tissue samples. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. The recent and fruitful literature on the latter topic in the context of microarrays is covered in depth. Distributional problems, including use of the t-distribution and its variants to provide robustness are introduced with a discussion of numerous methods, frequentist and Bayesian, to handle the multiplicity issue. The latter part of this chapter considers the clustering of genes that have been identified as being differentially expressed with a view to finding: (1) groups of genes that are significantly correlated with each other; (2) groups of genes that share similar expressions across the tissues.

Chapter 6 considers methods in discriminant analysis or supervised classification in a general context with a view to their application to microarray data. Discriminant

rules covered include the traditional normal-based linear and quadratic discriminant classifiers, more flexible parametric rules based on normal mixtures or mixtures of factor analyzers, support vector machines and their variants, nearest-neighbor and nearest centroid rules, classification trees, and neural networks. The problem of error-rate estimation of a discriminant rule is considered too, along with ways for the provision of standard errors for the estimates of the error rates.

Chapter 7 considers applications of some of the discriminant rules introduced in the previous chapter to the supervised classification of tissue samples. In applications concerned with the diagnosis of cancer, one class may correspond to cancer and the other to benign tumors. In applications concerned with patient survival following treatment for cancer, one class may correspond to the good prognosis group and the other to the poor prognosis group. Also, there is interest in the identification of "marker" genes that characterize the different tissue classes. Attention is focused on applications of the support vector machine and nearest-shrunken centroids, which is a recent version of nearest centroids to handle the very large number of genes. These two approaches are demonstrated on some cancer data sets. Particular attention is paid to the need to correct for the selection bias in estimating the prediction capacity of a discriminant rule formed from a subset of genes selected from a much larger set.

Chapter 8 is concerned with linking results of a model-based clustering of tumor tissues on cancer biology and clinical outcome. Cancer patients with the same stage of disease can have markedly different treatment responses and clinical outcome. Thus there is much interest in whether microarray expression data can be used to provide prognostic information beyond that provided by stage and other traditional clinical criteria. We report some recent results that show that the clustering provides significant prognostic information on the outcome of the disease beyond that available in current systems based on histopathology criteria and extent of disease at presentation.

The authors wish to acknowledge several Houston-based colleagues. LeeAnn Chastain contributed significant efforts to Chapter 1 with respect to the literature search, writing, editing, and proofreading. We are grateful to Wei Zhang, Keith Baggerly, Kevin Coombes, Li Zhang, and Tuyet-Trinh Do for reading and commenting of this chapter. Peter Müller was a great collaborator on the nonparametric Bayesian mixture model for differential gene expression. Sijin Wen assisted the second author in programming the gene shaving method. Bradley Broom contributed significantly with advanced programming using his high-performance computing tool.

Concerning acknowledgments to colleagues in Brisbane, the authors wish to thank Nazim Khan, Abdollah Khodkar, Katrina Monico, and Justin Zhu for their assistance. They wish to acknowledge the significant contributions made by Angus Ng and Liat Ben-Tovim Jones to the research leading to the results reported in Chapter 8. Further thanks are due to Liat for her many very helpful comments and suggestions on drafts of the manuscript. Finally, special thanks are due to Richard Bean who has greatly assisted with all aspects of the book, including proof reading and overseeing the numerous technical issues that arose during the preparation of the manuscript in camera-ready form.

The first author was supported by the Australian Research Council. The second author was partially supported by University of Texas SPORE in Prostate Cancer grant

CA90270, and the Early Detection Research Network grant CA99007. Thanks are due too to the authors and owners of copyright material for permission to reproduce tables and figures, and to Joel Tyndall (Institute for Molecular Bioscience) who used the InsightII Modeling Environment (Accelrys, 2004) to prepare the art work for the cover of the book.

Brisbane, Australia — Geoff McLachlan

Houston, USA — Kim-Anh Do

Compiègne, France — Christophe Ambroise