A Practical Approach to Microarray Data Analysis

Edited by Daniel P. Berrar, Werner Dubitzky and Martin Granzow

Kluwer Academic Publishers, Norwell, MA, USA; 2003; ISBN 1-4020-7260-0; €99.50, £65.00, US$99.50; Paperback; 368 pp.

This book comprises one introductory chapter and 19 technical chapters. It is aimed at 'life scientists, statisticians, computer experts, technology developers, managers and other professionals tasked with developing, deploying and using microarray technology'. The introductory chapter provides an informative overview of microarray analysis, from image processing through to biological validation. Some basic biology is provided for readers without a background in the life sciences. The technical chapters are thoughtfully presented in an order which reflects the overall data analysis process. The emphasis of the book is on the applied aspects of microarray analysis, but comprehensive literature cross-references direct the advanced reader to more theoretical work.

Chapters 2 to 6 deal with the pre-processing of data. Basic concepts are presented in Chapter 2 (Tinker and co-workers) for newcomers to microarray technology. The overview of data pre-processing provided by Chapter 2 is followed by four chapters addressing specific issues. Missing data, resulting from factors such as image corruption, insufficient resolution, or dust and scratches on the slide, is a problem encountered by all users of microarray technology. Moreover, many analysis algorithms (including PCA and SVD) require complete matrices to function. In Chapter 3 (Troyanskaya *et al.*), methods for missing value estimation are described, including their implementation using free publicly-available software tools. Systematic error and its removal using normalisation is the subject of Chapter 4 (Morrison and Hoyle). Popular methods of normalisation are reviewed with reference to both oligo- and spotted-array data.

One of the most significant challenges facing analysts of microarray experiments is the high-dimensional structure of the data, where the number of variables (genes) is much larger than the number of observations (samples). Typical microarray experiments may have thousands to tens-of-thousands of variables, but usually only tens of samples. If the goal of the analysis is to use a statistical test to determine which genes are differentially expressed, then the problem of multiple comparisons arises with an unprecedented severity. The high dimensionality of the data also impacts on machine learning techniques such as classification or clustering. Chapters 5 (Wall *et al.*) and 6 (Xing) tackle the 'curse of dimensionality', discussing a variety of classic and state-of-the-art approaches to feature selection and dimension reduction.

Having dealt with issues of data preparation, the book moves on to exploratory data analysis tasks. Such tasks can be broadly divided into two categories: predictive modelling (also known as supervised learning, classification or discriminant analysis); and pattern detection (often referred to as unsupervised learning, clustering or automatic class discovery). A typical application of predictive modelling would be the use of gene expression profiles to distinguish known tumour classes. Predictive modelling receives detailed discussion, with almost one-third of the chapters in the book devoted to this topic. The first of these, Chapter 7 (Dudoit and Fridlyand), outlines the statistical foundations of classification and describes two traditional classification approaches: linear discriminant analysis and nearest neighbour classifiers. Each of the subsequent chapters focuses on a specific, state-of-the-art methodology for supervised learning, namely Bayesian networks (Chapter 8 by Zhang and Hwang), support vector machines (Chapter 9 by Mukherjee), a weighted flexible compound covariate method with decision trees (Chapter 10 by Shyr and Kim), artificial neural networks (Chapter 11 by Ringnér *et al.*) and k-nearest neighbour and genetic algorithms (Chapter 12 by Li and Weinberg). The authors of Chapters 9, 11 and 12 helpfully recommend specific, free software tools for implementing the algorithms they describe. Additionally, Mukherjee provides practical guidelines for users of support vector machines and even suggests parameter settings for optimal performance.

In contrast to classification methodologies where the groups of interest are known prior to the analysis, pattern detection algorithms are concerned with screening the available data for unknown relationships. The goal of these methods is to identify groups of genes or samples that exhibit similar expression profiles. An example might be the identification of previously unrecognised, but clinically significant, subclasses of cancers. Chapter 13 introduces the fundamental aspects of clustering genomic expression data, including basic criteria for the selection of clustering techniques and assessing cluster validity. The following three chapters present a detailed discourse on a particular method for automatic class discovery. The application of hierarchical clustering methods to microarray data is the subject of Chapter 14 (Stanford *et al.*). Self-organising maps (SOMs) are described and critically evaluated in chapter 15 (Azuaje). Non-hierarchical, non-SOM methodologies, including the popular k-means algorithm, are discussed in Chapter 16 (Yeung). Practical guidelines are provided for the implementation of these methods, including recommendations for appropriate software.

Chapter 17 (Lin and Johnson) presents a gentle introduction to correlation and association analysis, which, unlike the majority of the chapters in this book, does not assume that the reader has anything more than an elementary understanding of mathematics. Given the relevance of correlation to much of the material in 'A Practical Approach to Microarray Analysis', this chapter might have been better placed towards the beginning, rather than the end of the book.

The ultimate aim of many microarray studies is to translate sets of differentially expressed genes into a functional profile, which will provide insight into the complex interactions that take place at the pathway level. In Chapter 18, Draghici and Krawetz examine the challenge of global functional profiling of gene expression data. The authors describe Onto-Express, a software package they have developed for this task and invite readers to use this system to analyse their own data.

The penultimate chapter (Leung *et al.*) provides a very useful review of both the free publicly-available software and commercial software packages for microarray data analysis. This chapter is complemented by an online software list (http://ihome.cuhk.edu.hk/∼b400559/arraysoft.html), which is an ongoing project of the authors. Those seeking analysis software, whether it be for primer design, image analysis, data mining, pathway reconstruction, annotation, laboratory information management systems (LIMS), or for a specific algorithm, would be well advised to consult this chapter. The book concludes with a discussion of microarray analysis from a process perspective, with particular reference to the practical issue of project management.

The editors are to be congratulated on compiling such a comprehensive, well organised and authoritative text on microarray analysis. Nevertheless, some clarification of the target audience for this book is needed. A sound background in mathematics and statistics is a prerequisite for understanding the majority for the chapters in this book. Most biologists are likely to find more than half of the chapters impenetrable. This is not to say there is nothing in this book for researchers who favour experimentation over calculation. Several of the chapters (notably 1, 7, 13, 17, 19 and 20) should be accessible to biologists with less statistical knowledge and will provide them with insight to a complex field. Overall, 'A Practical Approach to Microarray Analysis' represents an invaluable resource for statisticians, bioinformaticians and mathematically-talented biologists. For such readers, this book is perhaps the definitive guide to microarray analysis at present.

Matt Wayland
UK Human Genome Mapping Project Resource Centre
Cambridge, UK