Book Reviews

Statistics for Microarrays: Design, Analysis and Inference

Ernst Wit and John McClure
John Wiley & Sons Ltd, Chichester, UK, 2004; ISBN 0-470-84993-2; 278 pp.; Hardback; £45.00/€67.50

The aim of 'Statistics for Microarrays' is to explain the statistical methods commonly used for microarray analysis. The book is divided into two parts: the first, 'Getting Good Data', focuses on experimental design, normalisation and quality control issues. The second, 'Getting Good Answers', deals with higher level statistical inferences about the biological questions of interest, such as clustering samples or genes; assessing differential expression; and classification and prediction. The book opens with a useful section describing a number of experiments and associated datasets that are used throughout the book to illustrate the different stages in an analysis of a microarray dataset. An additional special feature is the inclusion of descriptions of R-language functions. Some of these are existing functions in standard R libraries, others are implementations by the authors of new methods developed in the book. The new functions are included in an R-package 'smida', which, along with the datasets, is made available online at the website for the book.

The book represents one of the first attempts to present a coherent exposition of the field of microarray data analysis, and is written in a clear and readable style. Although not a reference work, it has been written in such a way that one should not have to read the earlier chapters in order to understand the later ones. This idea that chapters or sections should be self-contained has been somewhat overdone and results in some statistical methods being explained more than once. For example, the definition of the t-statistic is given in both the classical and the Bayesian hypothesis sections in Chapter 8. The style is also rather repetitive within sections: for example, 'confounding' is explained well on page 37, yet a separate paragraph with the heading 'confounding' appears on page 38. In general, the use of cross-references within the book would have been helpful (eg Sammon plots are used at an early stage but are not introduced until later, with no cross-reference).

Chapters 3 and 4 on experimental design and normalisation are very good. There is a detailed discussion of the number of replicate arrays needed to detect a given level of fold change between experimental conditions; a discussion of the variability of pooled samples; and an extensive section on finding optimal designs for two-colour arrays. Normalisation is explained well, with plots illustrating the various reasons for normalisation. The discussion of single-channel arrays, in particular those of Affymetrix, is partly dealt with in separate subsections of Chapters 3 and 4.The resulting presentation is rather messy and also slightly misleading: the probes representing a gene are not technical replicates (as claimed in Chapter 3), but represent different subsequences of the sequence encoding a gene (as correctly stated in Chapter 4). The short description of some methods for estimating gene expression measures from oligonucleotide arrays at the end of the normalisation chapter does not do justice to the large literature on this subject.

The quality assessment chapter contains some instructive examples and good illustrations of the qualities of some of the most commonly used pairwise distance measures: the Euclidian, Manhattan and correlation distance measures. The chapter describes some interesting methods, such as Sammon mapping for dimensional reduction and 'false array images' for assessing array handling, but on the former point is not very clear. Sammon mapping is used to illustrate several different possible reasons for poor-quality data; however, how the different possible reasons would be distinguished is not obvious.

There are two chapters on clustering methods, one on unsupervised methods used to group samples and/or genes and one on supervised methods of classification. The chapter on unsupervised clustering starts with a good discussion of different possible measures for calculating distances between clusters. It focuses mainly on hierarchical and partitioning around medoids (PAM)-type algorithms, with a brief discussion of model-based clustering at the end. The authors rightly warn against putting too much faith in agglomerative hierarchical clustering of genes, although the point could have been made better with some illustrations of where this may go wrong, in line with the good illustrations in the book on the virtues of various distance measures.

The topic of gene-filtering is covered in the chapter on supervised methods. This chapter briefly introduces a number of important concepts in classification theory, predictor evaluation and cross-validation. Attention is restricted to simple methods, such as principal component analysis, linear discriminant analysis and k-nearest neighbour classification for class prediction and penalised and k-nearest neighbour regression for classifying and predicting continuous responses. Throughout the chapters on clustering, the authors present a number of new methods, particularly relating to the problem of selecting appropriate numbers of clusters or numbers of predictors, which are made available as R functions.

Differential gene expression is covered in a separate chapter. The standard varieties of t-test are described, along with guidance on which to use in various situations. Moreover, *p*-values and error rates (familywise and false discovery rates) are discussed and different methods of obtaining *p*-values (parametric, bootstrap and permutation) are given. This section provides a very good introduction to one of the most widely used methods for assessing differential expression. One

drawback is that there is no discussion of methods (such as the significance analysis of microarrays [SAM] method) for stabilising gene variance estimates used in t-statistics, which are often used when very few replicate arrays are available.

In summary, this book provides a good introduction to the statistical analysis of microarray data. The focus is primarily on cDNA arrays, although the higher-level analysis in the second half of the book can mostly be applied to oligonucleotide arrays as well. The more mathematically inclined reader may wish to refer to original papers for sophisticated discussion. The book should be well suited to the biologist or computer scientist who wants an overview of the problems encountered in analysing microarray data and to gain some understanding of the different methods available.

Anne-Mette Hein and Alex Lewin
Department of Epidemiology and Public Health
Imperial College, London
London, UK

Consanguinity, inbreeding, and genetic drift in Italy

Luigi-Luca Cavalli-Sforza, Antonio Moroni and Gianni Zei

Princeton University Press, Princeton, NJ, USA; 2004; ISBN: 0-691-08992-2 (Paperback); 0-691-08991-4 (Cloth); 315 pp.; £26.95/US$39.50 (Paperback); £51.95/$US 79.50 (Cloth)

This book tells the history of studies of classical polymorphisms, surnames and church records of marriages (and particularly dispensations to marry relatives) that began in the small communities of the Parma Valley during the 1950s and 1960s, and was later extended to the Italian islands and — in more limited form — the whole of Italy.

Although aspects of these studies have previously been published in several formats, this is the first full account of the background to the studies, the methods used, the results and the conclusions of the principal investigators. The publication of this book is therefore important, because these studies have formed a bedrock for late 20th century population genetics. Human population genetics has, and continues to be, marred by poorly designed sampling schemes. The careful design of these studies, together with statistical analyses and computer simulations that were often groundbreaking, remains an example to others.

Much of the book discusses consanguineous marriages — for example, marriages between cousins. The discussion covers Roman, 'German' (Lombard) and Catholic Church law, as well as other social, economic and demographic factors that affect the prevalence of such marriages. The effects of both consanguinity and 'random inbreeding' (geographically restricted mate choice) on genetic drift are also studied. There is also a chapter discussing their effects on both normal and pathological phenotypes.

In these days of genome-wide genetic surveys and fast computational analyses, the painstaking effort required to collect and analyse these relatively sparse data seems unthinkable. Although we can now answer the questions with much greater precision, the basic issues of genetic variation on a fine geographical scale — and its relationship to demographic factors, drift, selection and, ultimately, to phenotypes of interest — are the same as they were 50 years ago, when the studies described here were just beginning.

David Balding
Imperial College London
London, UK