Book Review

Design and Analysis of DNA Microarray Investigations

R. M. Simon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, Y. Zhao,
Springer-Verlag, New York, NY, USA; 2004;
ISBN: 0-387-00135-2; 199 pp.; Hardback;
58 greyscale illustrations and 15 colour illustrations; US$59.95

Microarray experiments have become standard methodology for measuring the expression levels of many genes across a range of cell lines, tissues or other organic material. The decidedly non-standard aspect of such experiments is the subsequent data analysis: the small number of noisy microarrays relative to the large number of genes presents a challenge to biologists using this technology. This book aims to aid the biologist in solving some of these problems by finding common ground in the methodologies that have been suggested thus far. It largely achieves this aim.

Like several other books about the analysis of microarray data, it is written by a collection of authors. Unlike its predecessors, it succeeds in making this fact invisible to the reader: the chapters are logically integrated and perfectly consistent with each other. Moreover, the book describes, overall, a sensible approach to microarray analysis. It was a relief to read the 'Class Comparison' chapter, which deals with finding differentially expressed genes. It describes hypothesis testing very sensibly and is refreshingly sceptical about methods that use only one replicate. Similarly, in the chapter on 'Microarray Design', the general discussion about replication is statistically sound. As part of the publication overflow on the topic of microarray analysis, all kinds of claims about replication have been doing the rounds, but Simon and co-authors are not distracted by them. For example, rather than recommending the rather wasteful dye-swap designs, they simply stress the importance of having dye balance across an experiment.

The book has both the advantage and the disadvantage of being relatively short. It gives biologists and bioinformaticists a quick overview of what is 'on the market', without excessive detail. In certain places, however, the book suffers from a lack of detail. In 'Class Comparisons', the reader has to search through dense text to find the different rejection rules for error control. The discussion on pooling samples is technically correct, but it covers only one page and might leave the unsuspecting biologist with the impression that 'the approach does not provide a valid basis for biological conclusions' (p. 16), whereas the opposite is true as long as each of the pools consists of biological replicates. The authors recommend against using loop designs (p. 21), but the problems mentioned can all be overcome, and this deserves further discussion. For example, it is not necessary at all that each sample should be hybridised twice within a loop design; in fact, such technical replication should be avoided.

In the chapter on class prediction, only the two-class prediction case is discussed. It is mentioned that this can be extended, but no details are offered on how that this can be achieved. In the same chapter, feature selection is described as a first, and separate, stage in building a classifier, in order to reduce the number of features (ie genes) under consideration (p. 97). Although mentioned again 13 pages later, the importance of incorporating this feature selection step into any classifier performance evaluation, such as cross-validation, does not receive sufficient attention: it is a common error in classification analyses.

In general, the book could have benefited from less dense explanations, brief 'take home' messages, a clearer layout and more examples. The colour inserts in the middle of the book do not really serve this purpose as well. It would have been better if more attention had been paid to drawing more adequate black-and-white images throughout the text itself.

I detected no outright errors in the book, but at a few points the authors are too uncritical. It is common practice in microarray analyses to exclude spots based on *ad hoc* rules, although it is almost always possible to include such spots by giving them low weights in analyses. The authors discuss a whole range of exclusion strategies. It is not clear to me why low-signal genes and, in particular, genes with low variance should be excluded at all (pp. 45–46). Moreover, since Affymetrix has altered its mismatch adjustment for the sole reason of making the expression values positive, the authors could have pointed out the unstatistical and *ad hoc* nature of this operation, rather than giving it serious attention (p. 37).

Despite a few glitches, which are to be expected in a topical book such as this, the book is very readable, and the authors have done an excellent job in presenting sensible statistical methodology to the biological community.

Ernst Wit
University of Glasgow
Glasgow, UK