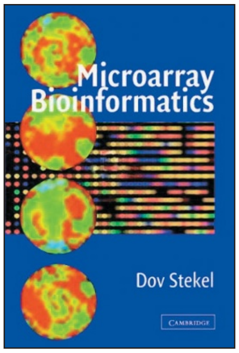
An array of problems



Microarray Bioinformatics

by Dov Stekel

Cambridge University Press, 2003
263 pp.; hardcover, $120; paperback,
$45 Cambridge University Press, 2003
ISBN 0521819822

Reviewed by Steven Russell

"To call in the statistician after the experiment is done may be no more than asking him to perform a post-mortem examination: he may be able to say what the experiment died of."

—R.A. Fisher, 1938

In the pre-genome era, molecular biologists discovered how genes function by a one man–one gene approach. Statistical data analysis was virtually unheard of, perhaps the odd Student's *t*-test, and experimental design involved seeking yes-or-no answers. In the post-genome era, molecular biologists became greedy: instead of studying one or a few genes at a time, why not look at them all? But surveying the expression of tens of thousands of genes simultaneously turns out not to be a simple matter of doing the experiment and publishing the paper: rather it is an analytical minefield, primed to confound innocent molecular biologists with a whole set of statistical difficulties.

Expression microarrays are conceptually simple; based on the exquisite specificity of nucleic acid hybridization, miniature arrays of unique DNA probes representing individual genes are interrogated with labeled RNA populations from cells or tissues of interest. Many experiments can be analyzed simultaneously to give a picture of how every gene in the genome is expressed in particular biological situations. Although this approach is potentially very powerful and may revolutionize how we view the regulation of gene expression, virtually every stage is a quagmire of variability and noisy data. So a successful microarray experiment must be conceived in the context of the methods to be used for subsequent analysis.

There are four key areas where biologists most need help: (i) the much-neglected area of data extraction, where computer software converts a scanned microarray image into a table of numerical data representing hybridization signal intensities for each probe; (ii) data normalization, where data is processed to enable comparisons between samples (either dual hybridizations on one array, or data from different arrays); (iii) the organization and selection of genes that are differentially expressed between samples; and (iv), designing sets of microarray hybridizations to maximize the statistical power of the analysis.

This book is structured like a course delivered by the author, a mathematician with considerable experience in the microarray field, to guide novices through all stages of microarray experiments while providing an understanding of why particular statistical treatments are necessary. Each chapter deals with a separate problem, then presents a set of worked examples using real data from the literature, followed by a summary and a good set of references. The first chapter introduces the state of the art in the production and use of microarrays. The chapter on probe annotation is probably superfluous for most researchers, who will usually know where to get the best sequence data for their system. The chapter on oligonucleotide design is biased towards the selection of short oligonucleotides; although it describes the concepts behind good design, it ignores much excellent published work on 50- to 70-nt probes and fails to direct readers to some relevant, first-class open-source software. The meat of the book covers the major issues that arise during data extraction and analysis. The weakest chapter is that on image analysis. Extracting numerical data from the scanned array image is a crucial step in the whole microarray process, and the old adage 'garbage in–garbage out' is particularly relevant. Current spot-finding and data extraction software is generally poor and, though the author hints at some of them, a far more comprehensive description of the difficulties in this area would help users appreciate that some of the software bundled with scanners is inadequate and that they should seek alternatives. The author does an excellent job of covering high-level analysis of microarray data, particularly clustering and partitioning. Also worthy of note is the chapter on normalization, which comprehensively describes the theory and techniques behind data normalization and includes recent methods for dealing with spatial variability. Similarly, the penultimate chapter, on experimental design, provides a helpful and clear guide to the bias and confounding effects that good design can help minimize; this chapter could usefully be expanded to hammer home the importance of sound design. The final chapter covers the need for standards and the MIAME community standard (now obligatory for microarray data submitted to this journal, among others).

Overall, this book makes an admirable attempt to cover a complex subject. It provides the statistically naive biologist with a gentle introduction to the data transformations and manipulations needed to deal with microarrays, and the worked examples with publicly available data are well described. Although more detail on image analysis and a less biased view of oligonucleotide design would improve it, I think the paperback version is excellent value for any budding arrayer, and will hopefully reduce the need for statistical coroners to hold embarrassing inquests about the quality of researchers' data. ■

Steven Russell is in the Department of Genetics at University of Cambridge, Downing Street, Cambridge CB2 3EH, UK. e-mail: s.russell@gen.cam.ac.uk