DNA Microarrays and Gene Expression: From Experiments to Data Analysis and Modeling
P. Baldi and G. W. Hatfield

Genomics has opened many new and exciting avenues of exploration for biologists. The potential now exists for developing an integrated molecular understanding of the genetic circuitries and networks underlying organismal development and physiology. DNA microarrays are perhaps the most important of the technologies springing up in the wake of genome sequencing. They are contributing to this revolution in biology and medicine by providing platforms for profiling the expression of thousands of genes simultaneously within individual cells, tissues or even whole organisms. The output from DNA array experiments can lead to the identification of individual genes that are differentially regulated between control and test states, and of whole gene sets that are co-regulated. DNA microarrays also hold great promise as diagnostic tools for evaluating an individual’s genetic risks for disease, for detecting the presence of disease and for clinical prognosis.

The book, written by Baldi and Hatfield, is an important and timely addition to the DNA microarray literature, although one underlying assumption is that most readers already have at least a cursory familiarity or direct interest in DNA array technology. Nonetheless, the first several chapters of the book provide an easy-to-digest overview of the current state of DNA microarrays. Included in these chapters are descriptions of various DNA array formats and platforms: photolithographic-generated oligonucleotide arrays (as popularised by Affymetrix Inc.), glass slide arrays and nylon arrays, to name but a few. This is followed by descriptions of instruments used for detecting and acquiring array readouts, such as laser confocal scanners used to detect Cy3 and Cy5 fluorophores and phosphorimagers capable of detecting radiisotopes.

Potential readers should realise, however, that this book is not intended to be a primer describing how to generate in-house arrays. As emphasised by the title, the book is primarily concerned with the topics of data acquisition, analysis and modelling.

As the book progresses, the authors’ experimental interest in Escherichia coli is revealed. Special attention is devoted to the methods and specific considerations associated with gene expression profiling in E. coli — datasets obtained using nylon arrays probed with ³²P-labelled cDNA targets feature prominently throughout the book. Of interest to many, the authors include a comparison of nylon filter data versus Affymetrix GeneChips™. Unfortunately for those relying on dual colour (Cy3/Cy5) DNA array readouts, a similar depth of coverage on this topic is not provided; readers must seek other sources to learn about issues such as dye-swaps and multi-factor analysis of variance (ANOVA).

The book shifts into high gear as it moves into the realm of statistical analysis and modelling of array data. The field of high-throughput genomics is responsible for generating a deluge of large datasets. It is universally agreed that methods for analysing these datasets are still evolving. It could even be argued that the most debated topic in the field of DNA arrays is how best to normalise and mine large
datasets in order to obtain ‘meaningful’ results. Thus, these chapters should be of particular interest to readers contemplating or already using DNA arrays. Baldi is no stranger when it comes to explaining the particulars of acquiring and handling large genomic datasets, having authored a highly regarded book, ‘Bioinformatics: the Machine Learning Approach’. In their book, Baldi and Hatfield cover a large number of topics including probabilistic modelling of array data, tools for visualising large datasets and algorithms for clustering data. Their treatment of these topics is not exhaustive, nor is it intended to be. Many investigators rely on commercial software packages for analysing and visualising DNA microarray data without necessarily having a clear understanding of the techniques being employed. These chapters thus provide a framework for understanding some of these concepts; however, those readers with a fairly solid foundation in statistics stand to gain the most benefit.

From a practical perspective, Chapter 7 can be considered one of the most useful chapters in the book because it takes the reader through the design, analysis and interpretation of gene profiling experiments using data obtained from *E. coli*. This chapter examines sources of errors arising from using different DNA arrays and mRNA samples and illustrates how one can estimate the global false-positive level for a given experiment. It is safe to assume that biologists and statisticians alike would agree that having a large number of experimental replicates can reduce error and increase the confidence one places in the identification of differentially expressed genes when the number of experimental replicates is low.

The final chapter of the book is devoted to a discussion of systems biology — how components such as DNA, RNA, proteins and other small molecules interact. The reader is reminded that DNA microarrays will play important roles in the integration of all of this information. Examples are provided, showing how attempts can be made to simulate various biological regulatory networks through the use of computational tools.

The book also contains numerous tables — providing useful information about DNA microarray equipment and software analysis packages — as well as four appendices. Appendix A contains a detailed set of DNA array hybridisation protocols that would serve as a good entry point for those contemplating using DNA arrays. Appendix B provides additional statistical complements to earlier chapters in the book. Appendix C lists a number of URLs that will connect readers to some of the more useful web-based microarray resources and forums. Included in this list is a URL for a web-based statistical package called CyberT developed, in part by Baldi, for analysing DNA array data. Appendix D provides additional information about CyberT and an illustration of the CyberT web interface.

Overall, the writing in this book is clear and concise. There is a cohesive flow from chapter to chapter because Baldi and Hatfield are the authors of the text, not simply the editors of a set of commissioned chapters. Each chapter ends with an extensive reference list. This book should be read by anyone interested in gaining an experimental perspective on the analysis, pitfalls and potentials of DNA array data.

Patricia E. Kuwabara
The Wellcome Trust Sanger Institute
Cambridge, UK