MICROARRAY BIOINFORMATICS

This book is a comprehensive guide to all of the mathematics, statistics and computing you will need to successfully operate DNA microarray experiments. It is written for researchers, clinicians, laboratory heads and managers, from both biology and bioinformatics backgrounds, who work with or who intend to work with microarrays. The book covers all aspects of microarray bioinformatics, giving you the tools to design arrays and experiments, to analyze your data, and to share your results with your organisation or with the international community. There are chapters covering sequence databases, oligonucleotide design, experimental design, image processing, normalisation, identifying differentially expressed genes, clustering, classification and data standards. The book is based on the highly successful Microarray Bioinformatics course at Oxford University and, therefore, is ideally suited for teaching the subject at the postgraduate or professional level.

Dov Stekel is Director of the Microarray Bioinformatics professional course at Oxford University and is a visiting academic at the Department of Biochemistry there. He also operates a bioinformatics consultancy, Bius, providing services to customers in the biotechnology industry and academia. Previously, he was the Manager of Bioinformatics at Ed Southern's microarray company, Oxford Gene Technology, and has worked as a Bioinformatics Scientist at Glaxo Wellcome. Dov lives in London and is an avid chocolate maker.
To my parents, Zsuzsi and Ronnie Stekel

In memory of Stephen Zatman, 1971–2002
Contents

Foreword  ix
Preface  xi
Acknowledgments  xiii

1  Microarrays: Making Them and UsingThem  1
2  Sequence Databases for Microarrays  19
3  Computer Design of Oligonucleotide Probes  43
4  Image Processing  62
5  Normalisation  73
6  Measuring and Quantifying Microarray Variability  100
7  Analysis of Differentially Expressed Genes  110
8  Analysis of Relationships Between Genes, Tissues or Treatments  139
9  Classification of Tissues and Samples  183
10  Experimental Design  211
11  Data Standards, Storage and Sharing  231

Appendix. MIAME Glossary  253
Index  259

Color section follows p. 128.
Foreword

DNA array technology is almost fifteen years old, and still rapidly evolving. It is one of very few platforms capable of matching the scale of sequence data produced by genome sequencing. Applications range from analysing single base changes, SNPs, to detecting deletion or amplification of large segments of the genome, CGH. At present, its most widespread use is in the analysis of gene expression levels. When carried out globally on all the genes of an organism, this analysis exposes its molecular anatomy with unprecedented clarity. In basic research, it reveals gene activities associated with biological processes and groups genes into networks of interconnected activities. There have been practical outcomes, too. Most notably, large-scale expression analysis has revealed genes associated with disease states, such as cancer, informed the design of new methods of diagnosis, and provided molecular targets for drug development.

At face value, the method is appealingly simple. An array is no more than a set of DNA reagents for measuring the amount of sequence counterparts among the mRNAs of a sample. However, the quality of the result is affected by several factors, including the quality of the array and the sample, the uniformity of hybridisation process, and the method of reading signals. Errors, inevitable at each stage, must be taken into account in the design of the experiment and in the interpretation of results. It is here that the scientist needs the help of advanced statistical tools.

Dr. Stekel is a mathematician with several years of experience in the microarray field. He has used his expertise in a company setting, developing advanced methods for probe design and for the analysis of large, complex data sets. This book is based on this practical experience and, more particularly, on experience gained in designing and running a course on Bioinformatics at the University of Oxford. The demand for this course showed that there are hundreds of biologists wanting to learn how to get the most from their microarray experiments. This book will help them to understand the nature of the data and the likely sources of error. It provides them with practical guidance and tools for handling large data sets and the statistical methods that can deal with them.

Ed Southern
Preface

DNA microarrays are devices that measure the expression of many thousands of genes in parallel. They have revolutionised molecular biology, and in the past five years their use has grown rapidly in academia, medicine, and the pharmaceutical, biotechnology, agrochemical and food industries.

One of the principal features of microarrays is the volume of quantitative data that they generate. As a result, the major challenge in the field is how to handle, interpret and make use of this data. The field of bioinformatics has come to mean the applications of mathematics, statistics and information technology in the biological sciences, and the bioinformatics of microarrays is the answer to that challenge.

This book is a comprehensive guide to all of the bioinformatics you will need to successfully operate DNA microarray experiments. It is written for researchers, clinicians, laboratory heads and managers, from both biology and bioinformatics backgrounds, who work with or who intend to work with microarrays. The book covers all aspects of microarray bioinformatics, giving you the tools to design arrays and experiments, to analyze your data, and to share your results with your organisation or with the international community. It has been inspired by the Microarray Bioinformatics professional course at Oxford University, and thus would also be suitable for teaching the subject at postgraduate or professional level.

The book assumes a minimum knowledge of molecular biology, computer use and statistics. On the biology front, readers will find it helpful if they have an understanding of the basic principles of molecular biology, i.e., DNA, RNA, transcription and translation, as well as the notions of genome sequencing and the existence of sequence databases. On the computing side, it is assumed that readers are familiar with the World Wide Web, and how to obtain data and software from it. No programming experience is needed to understand the book, although some of the ideas in the book would require programming skills to implement. It is also assumed that readers are familiar with the basic ideas of statistical descriptions of populations, such as means, standard deviations, histograms and scatterplots.

Where possible, the chapters include worked examples using real microarray data from published experiments. Each chapter ends with references to the data used, to a small selection of specialized research papers and textbooks relevant for further study, and to Internet resources and software relevant for the implementation of the methods described in it. Unpublished data sets and errata are available at the web site for this book, http://www.microarraybioinformatics.com.
Outline of Contents

The book is split into eleven chapters:

Chapter 1: Microarrays: Making Them and Using Them, gives an introduction to microarray technologies, the different platforms by which microarrays are manufactured, and the laboratory process involved in microarray use.

Chapter 2: Sequence Databases for Microarrays, is a description of the international sequence databases that are used for microarray design and annotation.

Chapter 3: Computer Design of Oligonucleotide Probes, describes the algorithmic methods by which oligonucleotide probes for microarrays can be designed.

Chapter 4: Image Processing, looks at the computational algorithms used to convert microarray images into quantitative data.

Chapter 5: Normalisation, describes methods that are used to eliminate systematic bias introduced by the microarray platform from microarray data.

Chapter 6: Measuring and Quantifying Microarray Variability, describes methods for measuring and quantifying the stratified variability that is a feature of microarray data.

Chapter 7: Analysis of Differentially Expressed Genes, looks at the analysis of microarray data where the microarray is being used to identify genes that may be up-regulated or down-regulated in different tissues or conditions.

Chapter 8: Analysis of Relationships Between Genes, Tissues or Treatments, describes methods that are used to explore the relationships between different genes or samples, including clustering and other related methods.

Chapter 9: Classification of Tissues and Samples, discusses methods that can be used to build predictive models that use gene expression for diagnostic or prognostic purposes.

Chapter 10: Experimental Design, looks at a number of issues in relation to how to design a microarray experiment, including how to determine the number of replicates you would need to use.

Chapter 11: Data Standards, Storage and Sharing, describes the computer technologies needed to run a microarray laboratory, and the standards by which microarray experiments and data can be annotated and shared.
Acknowledgments

Who is wise? One who can learn from everybody.
– Shimon son of Zoma, 1st Century CE

My whole life has been extremely simple. For the first half I sat facing a blackboard and
for the second half I stood back to a blackboard. With regard to a blackboard I have made
only one complete turn – and with that my biography is complete.
– Nishida Kitaro, 1870–1945

This is the first book I have written, and it is the most difficult project I have taken
on in my life. It would not have been possible to start the book, let alone complete
it, without the inspiration, guidance, help, advice and support of my family, friends,
colleagues and students.

Ed Southern has supported this book in every possible way: personally, with friend-
ship and encouragement; scientifically, by giving me access to his slides that formed
the basis of Chapter 1, and by providing feedback on other chapters; and financially,
via access to an MRC research fellowship. It is a rare privilege to work with such an
outstanding individual, and one that I deeply treasure.

The inspiration for this book has been the Microarray Bioinformatics course that
is run at the Department of Continuing Education at Oxford University, and at the
Roslin Institute in Edinburgh. Every chapter of this book contains knowledge and
understanding that I have taken from the course and its lecturers; thank you to Andy
Greenfield, Clare Pritchard, David Wild, Ed Southern, Francesco Falciani, Kate Rice,
Lorenz Wernisch, Louise Pedlar, Pete Underhill, Rob Andrews and Sarah Webb. The
course would not be possible without the administrative support of Roni McGowan
and Rachel Cox, who have also been deeply helpful and supportive of this book. To
this date, more than 100 delegates have come through the course; I have learned from
each and every one of you, though you are too numerous to list in this short section.

Many figures in this book have been reproduced from lectures given on the mi-
croarray course. To have had access to such excellent lecture slides and figures has
been a tremendous bonus for which I am deeply appreciative. I thank Oxford Univer-
sity for permission to use these figures, and specifically Clare Pritchard (Figures 1.1a
and 1.10), Ed Southern (Figures 1.1b, 1.1c, 1.3, 1.4, 1.5, 1.6, 1.7 and 1.9), Francesco
Falciani (Figures 1.1d and 1.11), Rob Andrews (Figure 11.3), Roni McGowan (Figures
1.1a and 1.10) and Sarah Webb (Figures 1.12, 1.13, 1.14, 4.2, 4.4 and 4.5). Pete Underhill
helped produce Figures 4.1a, 4.3 and 4.5 at the MRC Mouse Genetics Unit in Harwell.
ACKNOWLEDGMENTS

Figure 1.4 is reproduced with permission from Affymetrix, and Figure 11.3 with permission from the Gene Ontology (GO) Consortium.

I have also been very fortunate to have been able to call on so many friends and colleagues to read chapters and provide feedback on them. Thank you to Ann Git, Chris Littler, Claire Parker, Clare Pritchard, David Wild, Ed Southern, Pete Underhill, Richard Talbot, Rob Andrews, Sarah Webb and Yoav Git, without whom this book would contain many more errors and inconsistencies. Any remaining mistakes are mine alone.

Anyone who has written a book knows that it is a long and lonely process. Nicky Press was a constant source of strength, support and companionship, not to mention hot lunches and cups of tea when I have most needed them. My writer friends: Daniel Litvin, Margaret Myers, Samantha Ellis and Sandra Shulman have a particular understanding of the exercise, and I thank them for their support, encouragement and advice.

Finally, and most importantly, I would like to thank my parents and sister for their unconditional love and support.
MICROARRAY BIOINFORMATICS