DNA MICROARRAYS AND GENE EXPRESSION

From experiments to data analysis and modeling

Massive data acquisition technologies, such as genome sequencing, high-throughput drug screening, and DNA arrays are in the process of revolutionizing biology and medicine. Using the mRNA of a given cell, at a given time, under a given set of conditions, DNA microarrays can provide a snapshot of the level of expression of all the genes in the cell. Such snapshots can be used to study fundamental biological phenomena such as development or evolution, to determine the function of new genes, to infer the role that individual genes or group of genes may play in diseases, and to monitor the effect of drugs and other compounds on gene expression. This interdisciplinary introduction to DNA arrays will be essential reading for researchers wanting to take advantage of this powerful new technology.

PIERRE BALDI is Professor and Director of the Institute for Genomics and Bioinformatics in the Department of Information and Computer Science and in the Department of Biological Chemistry at the University of California, Irvine.

WES HATFIELD is a Professor in the Department of Microbiology and Molecular Genetics in the College of Medicine and the Department of Chemical Engineering and Material Sciences in the School of Engineering at the University of California, Irvine.

DNA MICROARRAYS AND GENE EXPRESSION

From experiments to data analysis and modeling

PIERRE BALDI

University of California, Irvine

and

G. WESLEY HATFIELD

University of California, Irvine

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9780521176354

First published 2002
Reprinted 2003
First paperback edition 2011

A catalogue record for this publication is available from the British Library

Library of Congress Cataloguing in Publication data

Baldi, Pierre.
DNA microarrays and gene expression / Pierre Baldi and G. Wesley Hatfield.
p. cm.
Includes bibliographical references and index.
ISBN 0 521 80022 6
1. DNA microarrays. 2. Gene expression. I. Hatfield, G. Wesley, 1940– II. Title.
QP624.5.D726 .B353 2002
572.8′65 – dc21 2001052862

ISBN 978-0-521-80022-8 Hardback
ISBN 978-0-521-17635-4 Paperback

Additional resources for this publication at www.cambridge.org/9780521176354

Contents

Preface

A number of array-based technologies have been developed over the last several years, and technological development in this area is likely to continue at a brisk pace. These technologies include DNA, protein, and combinatorial chemistry arrays. So far, DNA arrays designed to determine gene expression levels in living cells have received the most attention. Since DNA arrays allow simultaneous measurements of thousands of interactions between mRNA-derived target molecules and genome-derived probes, they are rapidly producing enormous amounts of raw data never before encountered by biologists. The bioinformatics solutions to problems associated with the analysis of data on this scale are a major current challenge.

Like the invention of the microscope a few centuries ago, DNA arrays hold promise of transforming biomedical sciences by providing new vistas of complex biological systems. At the most basic level, DNA arrays provide a snapshot of all of the genes expressed in a cell at a given time. Therefore, since gene expression is the fundamental link between genotype and phenotype, DNA arrays are bound to play a major role in our understanding of biological processes and systems ranging from gene regulation, to development, to evolution, and to disease from simple to complex. For instance, DNA arrays should play a role in helping us to understand such difficult problems as how each of us develops from a single cell into a gigantic supercomputer of roughly 10^{15} cells, and why some cells proliferate in an uncontrolled manner to cause cancer.

One notable difference between modern DNA array technology and the seventeenth-century microscope, however, is in the output produced by these technologies. In both cases, it is an image. But unlike the image one sees through a microscope, an array image is not interpretable by the human eye. Instead, each individual feature of the DNA array image must be measured and stored in a large spreadsheet of numbers with tens to

tens-of-thousands of rows associated with gene probes, and as many columns or experimental conditions as the experimenter is willing to collect. As a side note, this may change in the future and one could envision simple diagnostic arrays that can be read directly by a physician.

Clearly, the scale and tools of biological research are changing. The storage, retrieval, interpretation, and integration of large volumes of data generated by DNA arrays and other high-throughput technologies, such as genome sequencing and mass spectrometry, demand increasing reliance on computers and evolving computational methods. In turn, these demands are effecting fundamental changes in how research is done in the life sciences and the culture of the biological research community. It is becoming increasingly important for individuals from both life and computational sciences to work together as integrated research teams and to train future scientists with interdisciplinary skills. It is inevitable that as we enter farther into the genomics era, single-investigator research projects typical of research funding programs in the biological sciences will become less prevalent, giving way to more interdisciplinary approaches to complex biological questions conducted by multiple investigators in complementary fields. Statistical methods, in particular, are essential for the interpretation of high-throughput genomic data. Statistics is no longer a poor province of mathematics. It is rapidly becoming recognized as the central language of sciences that deal with large amounts of data, and rely on inferences in an uncertain environment.

As genomic technologies and sequencing projects continue to advance, more and more emphasis is being placed on data analysis. For example, the identification of the function of a gene or protein depends on many things including structure, expression levels, cellular localization, and functional neighbors in a biochemical pathway that are often co-regulated and/or found in neighboring regions along the chromosome. Clearly then, establishing the function of new genes can no longer depend on sequence analysis alone but requires taking into account additional sources of information including phylogeny, environment, molecular and genomic structure, and metabolic and regulatory networks. By contributing to the understanding of these networks, DNA arrays already are playing a significant role in the annotation of gene function, a fundamental task of the genomics era. At the same time, array data must be integrated with sequence data, with structure and function data, with pathway data, with phenotypic and clinical data, and so forth. New biological discoveries will depend strongly on our ability to combine and correlate these diverse data sets along multiple dimensions and scales. Basic research in bioinformatics

must deal with these issues of systems and integrative biology in a situation where the amount of data is growing exponentially.

As these challenges are met, and as DNA array technologies progress, incredible new insights will surely follow. One of the most striking results of the Human Genome Project is that humans probably have only on the order of twice the number of genes of other metazoan organisms such as the fly. While these numbers are still being revised, it is clear that biological complexity does not come from sheer gene number but from other sources. For instance, the number of gene products and their interactions can be greatly amplified by mechanisms such as alternative mRNA splicing, RNA editing, and post-translational protein modifications. On top of this, additional levels of complexity are generated by genetic and biochemical networks responsible for the integration of multiple biological processes as well as the effects of the environment on living cells. Surely, DNA and protein array technologies will contribute to the unraveling of these complex interactions. At a time when human cloning and organ regeneration from stem cells are on the horizon, arrays should help us to further understand the old but still largely unanswered questions of nature versus nurture and perhaps strike a new balance between the reductionist determinism of molecular biology and the role of chance, epigenetic regulation, and environment on living systems. However, while arrays and other high-throughput technologies will provide the data, new bioinformatics innovations must provide the methods for the elucidation of these complex interactions.

As we progress into the genomics era, it is anticipated that DNA array technologies will assume an increasing role in the investigation of evolution. For example, DNA array studies could shed light on mechanisms of evolution directly by the study of mRNA levels in organisms that have fast generation times and indirectly by giving us a better understanding of regulatory circuits and their structure, especially developmental regulatory circuits. These studies are particularly important for understanding evolution for two obvious reasons: first, genetic adaptation is very constrained since most non-neutral mutations are disadvantageous; and second, simple genetic changes can serve as "amplifiers" in the sense that they can produce large developmental changes, for instance doubling the number of wings in a fly.

On the medical side, DNA arrays ought to help us better understand complex issues concerning human health and disease. Among other things, they should help us tease out the effects of environment and life style, including drugs and nutrition, and help usher in the individualized molecular medicine of the future. For example, daily doses of vitamin C

recommended in the literature vary over three orders of magnitude. In fact, the optimal dose for all nutritional supplements is unknown. Information from DNA array studies should help define and quantify the impact of these supplements on human health. Furthermore, although we are accustomed to the expression "the human body", large response variabilities among individuals due to genetic and environmental differences are observed. In time, the information obtained from DNA array studies should help us tailor nutritional intake and therapeutic drug doses to the makeup of each individual.

Throughout the second half of the twentieth century, molecular biologists have predominantly concentrated on single-gene/single-protein studies. Indeed, this obsessive focus on working with only one variable at a time while suppressing all others in *in vitro* systems has been a hallmark of molecular biology and the foundation for much of its success. As we enter into the genomics era this basic paradigm is shifting from the study of single-variable systems to the study of complex interactions. During this same period, cell biologists have been following mRNA and/or protein levels during development, and much of what we know about development has been gathered with techniques like *in situ* hybridization that have allowed us to define gene regulatory mechanisms and to follow the expression of individual genes in multiple tissues. DNA arrays give us the additional ability to follow the expression levels of all of the genes in the cells of a given tissue at a given time.

As old and new technologies join forces, and as computational scientists and biologists embrace the high-throughput technologies of the genomics era, the trend will be increasingly towards a systems biology approach that simultaneously studies tens of thousands of genes in multiple tissues under a myriad of experimental conditions. The goal of this systems biology approach is to understand systems of ever-increasing complexity ranging from intracellular gene and protein networks, to tissue and organ systems, to the dynamics of interactions between individuals, populations, and their environments. This large-scale, high-throughput, interdisciplinary approach enabled by genomic technologies is rapidly becoming a driving force of biomedical research particularly apparent in the biotechnology and pharmaceutical industries. However, while the DNA array will be an important workhorse for the attainment of these goals, it should be emphasized that DNA array technology is still at an early stage of development. It is cluttered with heterogeneous technologies and data formats as well as basic issues of noise, fidelity, calibration, and statistical significance that are still being sorted out. Until these issues are resolved

and standardized, it will not be possible to define the complete genetic regulatory network of even a well-studied prokaryotic cell. In the meantime, most progress will continue to come from focused incremental studies that look at specific networks and specific interacting sets of genes and proteins in simple model organisms, such as the bacterium *Escherichia coli* or the yeast *Saccharomyces cerevisiae*.

In short, the promise of DNA arrays is to help us untangle the extremely complex web of relationships among genotypes, phenotypes development, environment, and evolution. On the medical side, DNA arrays ought to help us understand disease, create new diagnostic tools, and help usher in the individualized molecular medicine of the future. DNA array technology is here and progressing at a rapid pace. The bioinformatics methods to process, analyze, interpret, and integrate the enormous volumes of data to be generated by this technology are coming.

Audience and prerequisites

In 1996, Hillary Rodham Clinton published a book titled *It Takes a Village*. This book discusses the joint responsibilities of different segments of a community for raising a child. Just like it takes individuals with different talents to raise a child "it takes a village" to do genomics. More precisely, it takes an ongoing dialog and a two-way flow of information and ideas between biologists and computational scientists to develop the designs and methods of genomic experiments that render them amenable to rigorous analysis and interpretation. This book seeks to foster these interdisciplinary interactions by providing in-depth descriptions of DNA microarray technologies that will provide the information necessary for the design and execution of DNA microarray experiments that address biological questions of specific interest. At the same time, it provides the details and discussions of the computational methods appropriate for the analysis of DNA microarray data. In this way, it is anticipated that the computational scientists will benefit from learning experimental details and methods, and that the biologist will benefit from the discussions of the methods for the analysis and interpretation of data that results from these high dimensional experiments.

At this time, most biologists depend on their computational colleagues for the development of data analysis methods and the computational scientists depend upon their biologist colleagues to perform experiments that address important biological questions and to generate data. Since this book is directed to both fields, we hope that it will serve as a catalyst to facilitate these critical interactions among researchers of differing talents and

expertise. In other words, we have tried to write a book for all members of the genomic village. In this effort, we anticipate that a biologist's full appreciation of the more computationally intense sections might require consultation with a computational colleague, and that the computational scientists will benefit from discussions concerning experimental details and strategies with the biologist. For the most part, however, this book should be generally comprehensible by either a biologist or a computational scientist with a basic background in biology and mathematics. It is written for students, mostly at the graduate but possibly at the undergraduate level, as well as academic and industry researchers with a diverse background along a broad spectrum from computational to biomedical sciences. It is perhaps fair to say that the primary readers we have in mind are researchers who wish to carry out and interpret DNA array experiments. To this end, we have endeavored to provide succinct explanations of core concepts and techniques.

Content and general outline of the book

We have tried to write a comprehensive but reasonably concise introductory book that is self-contained and gives the reader a good sense of what is available and feasible today. We have not attempted to provide detailed information about all aspects of arrays. For example, we do not describe how to build your own array since this information can be obtained from many other sources, ranging from Patrick Brown's web site at Stanford University to a book by Schena *et al*.[1] Instead, we focus on DNA array experiments, how to plan and execute them, how to analyze the results, what they are good for, and pitfalls the researcher may encounter along the way.

The topics of this book reflect our personal biases and experiences. A significant portion of the book is built on material from articles we have written, our unpublished observations, and talks and tutorials we have presented at several conferences and workshops. While we have tried to quote relevant literature, we have concentrated our main effort on presenting the basic concepts and techniques and illustrating them with examples. The main focus of this book is on methods – how to design, execute and interpret a gene expression profiling experiment in a way that remains flexible and open to future developments.

In Chapter 1 we present a brief history of genomics that traces some of

[1] Brown, P. http://cmgm.stanford.edu/pbrown; Schena, M. (ed.) *Microarray Biochip Technology*. 2000. Eaton Publishing Co., Natick, MA.

the milestones over the past 50 years or so that have ushered us into the genomics era. This history emphasizes the technological breakthroughs – and the authors' bias towards the importance of the model organism *Escherichia coli* in the development of the paradigms of modern molecular biology – that have led us from the enzyme period to the genomics era.

In Chapter 2 we describe the various DNA array technologies that are available today. These technologies range from *in situ* synthesized arrays such as the Affymetrix GeneChip™, to pre-synthesized nylon membrane and glass slide arrays, to newer technologies such as electronic and bead-based arrays.

In Chapter 3 we describe the methods, technology, and instrumentation required for the acquisition of data from DNA arrays hybridized with radioactive-labeled or fluorescent-labeled targets.

In Chapter 4 we consider issues important for the design and execution of a DNA array experiment with special emphasis on problems and pitfalls encountered in gene expression profiling experiments. Special consideration is given to experimental strategies to deal with these problems and methods to reduce experimental and biological sources of variance.

In Chapter 5 we deal with the first level of statistical analysis of DNA array data for the identification of differentially expressed genes. Due to the large number of measurements from a single experiment, high levels of noise, and experimental and biological variabilities, array data is best modeled and analyzed using a probabilistic framework. Here we review several approaches and develop a practical Bayesian statistical framework to effectively address these problems to infer gene changes. This framework is applied to experimental examples in Chapter 7.

In Chapter 6 we move to the next level of statistical analysis involving the application of visualization, dimensionality reduction, and clustering methods to DNA array data. The most popular dimensionality and clustering methods and their advantages and disadvantages are surveyed. We also examine methods to leverage array data to identify DNA genomic sequences important for gene regulation and function. Mathematical details for Chapters 5 and 6 are presented in Appendix B.

In Chapter 7 we present a brief survey of current DNA array applications and lead the reader through a gene expression profiling experiment taken from our own work using pre-synthesized (nylon membrane) and *in situ* synthesized (Affymetrix GeneChip™) DNA arrays. Here we describe the use of software tools that apply the statistical methods described in Chapters 5 and 6 to analyze and interpret DNA array data. Special emphasis is given to methods to determine the magnitude and sources of experi-

mental errors and how to use this information to determine global false positive rates and confidence levels.

Chapter 8 covers several aspects of what is coming to be known as systems biology. It provides an overview of regulatory, metabolic, and signaling networks, and the mathematical and software tools that can be used for their investigation with an emphasis on the inference and modeling of gene regulatory networks.

The appendices include explicit technical information regarding: (A) protocols, such as RNA preparation and target labeling methods, for DNA array experiments; (B) additional mathematical details about, for instance, support vector machines; (C) a section with a brief overview of current database resources and other information that are publicly available over the Internet, together with a list of useful web sites; and (D) an introduction to CyberT, an online program for the statistical analysis of DNA array data.

Finally, a word on terminology. Throughout the book we have used for the most part the word "array" instead of "microarray" for two basic reasons: first, in our minds DNA arrays encompass DNA microarrays; second, at what feature density or physical size an array becomes a microarray is not clear. Also, the terms "probe" and "target" have appeared interchangeably in the literature. Here we keep to the nomenclature for probes and targets of northern blots familiar to molecular biologists; we refer to the nucleic acid attached to the array substrate as the "probe" and the free nucleic acid as the "target".

Acknowledgements

Many colleagues have provided us with input, help, and support. At the risk of omitting many of them, we would like to thank in particular Suzanne B. Sandmeyer who has been instrumental in developing a comprehensive program in genomics and the DNA Array Core Facility at UCI, and who has contributed in many ways to this work. We would like to acknowledge our many colleagues from the UCI Functional Genomics Group for the tutelage they have given us at their Tuesday morning meetings, in particular: Stuart Arfin, Lee Bardwell, J. David Fruman, Steven Hampson, Denis Heck, Dennis Kibler, Richard Lathrop, Anthony Long, Harry Mangalam, Calvin McLaughlin, Ming Tan, Leslie Thompson, Mark Vawter, and Sara Winokur. Outside of UCI, we would like to acknowledge our collaborators Craig J. Benham, Rob Gunsalus, and David Low. We would like also to thank Wolfgang Banzhaf, Hamid Bolouri, Hidde de Jong, Eric Mjolsness,

James Nowick, and Padhraic Smyth for helpful discussions. Hamid and Hidde also provided graphical materials, James and Padhraic provided feedback on an early version of Chapter 8, and Wolfgang helped proofread the final version. Additional material was kindly provided by John Weinstein and colleagues, and by Affymetrix including the image used for the cover of the book. We would like to thank our excellent graduate students Lorenzo Tolleri, Pierre-François Baisnée, and Gianluca Pollastri, and especially She-pin Hung, for their help and important contributions. We gratefully acknowledge support from Sun Microsystems, the Howard Hughes Medical Institute, the UCI Chao Comprehensive Cancer Center, the UCI Institute for Genomics and Bioinformatics (IGB), the National Science Foundation, the National Institutes of Health, a Laurel Wilkening Faculty Innovation Award, and the UCI campus administration, as well as a GAANN (Graduate Assistantships in Areas of National Need Program) and a UCI BREP (Biotechnology Research and Education Program) training grant in Functional and Computational Genomics. Ann Marie Walker at the IGB helped us with the last stages of this project. We would like also to thank our editors, Katrina Halliday and David Tranah at Cambridge University Press, especially for their patience and encouragement, and all the staff at CUP who have provided outstanding editorial help. And last but not least, we wish to acknowledge the support of our friends and families.