Richard M. Simon
Lisa M. McShane
George W. Wright
Edward L. Korn
Michael D. Radmacher
Yingdong Zhao

Design and Analysis of DNA Microarray Investigations

With 58 Figures, 15 in Full color

Springer

Richard M. Simon
Edward L. Korn
Lisa M. McShane
George W. Wright
Yingdong Zhao
Biometric Research Branch
National Cancer Institute
9000 Rockville Pike
MSC 7434
Bethesda, MD 20892-7434

Michael D. Radmacher
Departments of Mathematics
& Biology
Kenyon College
Gambier, OH 43022

Library of Congress Cataloging-in-Publication Data
Design and analysis of DNA microarray investigations / Richard M. Simon . . . [et al.].
p. cm. — (Statistics for biology and health)
ISBN 0-387-00135-2 (hbk. : alk. paper)
1. DNA microarrays—Statistical methods. I. Simon, Richard M., 1943– II. Series.
QP624.5.D726D475 2003
572.8'65—dc21 2003054790

ISBN 0-387-00135-2 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10898178

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

Statistics for Biology and Health

Series Editors
K. Dietz, M. Gail, K. Krickeberg, J. Samet, A. Tsiatis

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

Acknowledgments

We thank our colleagues at the National Cancer Institute who have given us the opportunity to become involved in cancer genomics, and to contribute to discovery of a new generation of therapeutics based on improved knowledge of tumor biology. We particularly thank Dr. Robert Wittes for supporting the establishment of the Molecular Statistics and Bioinformatics Section of the Biometric Research Branch and providing it with the independence to develop expertise, conduct independent research, and establish a unique multidisciplinary environment in which methodologists and experimentalists can interact.

We thank Amy Peng Lam for her development of BRB-ArrayTools in collaboration with Richard Simon, and for her many contributions to our microarray analyses. Thanks to Dr. Ming-Chung Li for excellent statistical computing support, to Dr. James P. Brody for permission to use Figure 3.1, to the National Human Genome Research Institute for illustrations from their Talking Glossary website, to Drs. Tatiana Dracheva, Jin Jen, and Joanna Shih for providing Affymetrix GeneChip™ images, to Mr. Erik Marchese at Affymetrix for technical consultation, and to Dr. Laura Lee Johnson for a careful reading of an earlier draft.

Contents

1

Introduction

DNA microarrays are an important technology for studying gene expression. With a single hybridization, the level of expression of thousands of genes, or even an entire genome, can be estimated for a sample of cells. Consequently, many laboratories are attempting to utilize DNA microarrays in their research. Whereas laboratories are well prepared to address the significant experimental challenges in obtaining reproducible data from this RNA-based assay, investigators are less prepared to analyze the large volumes of data produced by DNA microarrays.

Although many software packages have been developed for the analysis of DNA microarray data, software alone is insufficient. One needs knowledge about the various aspects of data analysis in order to select and utilize software effectively. There is a plethora of analysis methods being published and it is difficult for biologists to determine which methods are valid and appropriate for their problems.

Many scientists have learned that software is not an adequate substitute for biostatistical knowledge and seek statistical collaborators. Unfortunately, there is presently a shortage of statisticians who are available and knowledgeable about DNA microarrays. For statisticians to be effective collaborators in any area, they must invest the time to understand the subject matter area and become familiar with the literature so that they can ask the right questions and identify the key issues.

Our objectives in this book are twofold: to provide scientists with information about the design and analysis of studies using DNA microarrays that will enable them to plan and analyze their own studies or to work with statistical collaborators effectively, and to aid statistical and computational scientists wishing to develop expertise in this area.

We believe that the design and analysis of microarray studies should be driven by the objectives of the experiment. We have identified several common types of objectives and for each type we have presented methods that we believe are statistically sound and effective. These methods are described in a manner that we believe will be understandable to most scientists. We empha-

size the concepts behind the methods rather than the mechanics of the use of the formulas. In most cases, the methods are available in existing software and the investigator will need knowledge of concepts to select methods and software more than knowledge of formulas for doing the calculations by hand. We have made the data used as examples in this book available on our Web site (see Appendix B) and have provided readers with a tutorial on the use of our BRB-ArrayTools software for analysis of these datasets. BRB-ArrayTools, described in Appendix C, is a menu-driven program incorporating many advanced analysis features but easily usable by scientists. BRB-ArrayTools is available without charge for noncommercial purposes.

We have tried to keep each chapter focused and relatively short in order to enhance its readability. Analytic methods for DNA microarray data are an active area of research. We have presented specific methods that we have found to be valid and useful. Although we generally describe a variety of approaches to analysis, we have not tried to be encyclopedic with regard to the literature. We hope that this serves the needs of most scientists looking for expert advice about sound and effective methods and also the needs of statistical and computational scientists looking for a broader coverage of the literature. Most of the material included has been written to be understandable to biological scientists without substantial statistical training. We have avoided mathematical and statistical derivations and nonessential notation.

Chapter 2 is a brief description of microarray platforms commonly used for gene expression profiling, including dual-label cDNA and oligonucleotide platforms and Affymetrix GeneChip™ arrays. Some experimentalists may choose to skip this chapter. Although microarrays can also be used for purposes other than gene expression profiling, such as sequencing and genotyping, these latter applications are not the focus of this book.

Chapter 3 discusses important aspects of the design of studies that use DNA microarrays. A complete presentation of the area of biomedical study design is not possible in one chapter, but we attempt to address many topics of special relevance in DNA microarray based studies.

Chapter 4 addresses the creation and analysis of images of intensities on microarrays after hybridization of labeled targets to the immobilized probes. That is, we discuss how pixel-level data are converted to probe-level or gene-level summaries. Although scientists generally do not do their own image analysis, some need to select software and to evaluate their images. Hence a basic understanding of the issues involved is useful.

Chapters 5 and 6 examine a variety of signal-processing issues, which must be addressed before the objective-directed analysis strategy is implemented. Chapter 5 covers methods of evaluating quality of microarray data. These issues are discussed separately for dual-label arrays and for GeneChips™. Chapter 6 addresses issues of normalization. Normalization is necessary because the raw intensities of labeled targets vary among arrays due to sources of experimental variability independent of level of expression. The objec-

tives of normalization are somewhat different for dual-label arrays and for GeneChips™, and both are discussed.

Chapters 7 through 9 present analysis strategies for studies where the major objectives are class comparison, class prediction, and class discovery, respectively. In class comparison problems discussed in Chapter 7 there is a predefined classification of the specimens and the objective is usually to determine which genes are differentially expressed among the classes. For example, comparing expression profiles for different types of tissue or for the same tissue under different conditions are class comparison objectives.

In some studies, particularly those involving expression profiles of diseased human tissues, there are predefined classes and the emphasis is on attempting to develop a gene expression-based predictor of the class to which a new specimen belongs. Such class prediction problems, and the related problem of prognostic prediction, are addressed in Chapter 8. For example, we may have tissues from patients with a specified disease who have received a specific treatment. One class may be those specimens from patients who responded to the treatment and the other class may be those tissues from patients who did not respond. The objective may be to predict whether a new patient is likely to respond based on the expression profile of his or her tissue specimen. Accurate prediction is of obvious value in treatment selection. In Chapter 8 we discuss the key components of a class prediction algorithm and describe several commonly used methods of prediction.

Chapter 9 addresses class discovery objectives. This includes discovery of new groupings or taxonomies of the specimens, based on expression profiles. Discovering classes of coexpressed and potentially coregulated genes is also a discovery objective. Class discovery is usually adressed using methods of cluster analysis. Chapter 9 also describes principal components analysis and multidimensional scaling and the graphical displays associated with these methods.

We present the material in Chapters 7 through 9 in a relatively nonmathematical style that will be understandable to a broad range of scientists and to illustrate many of the methods with examples.

Appendix A provides basic information on the biology of gene expression for statistical and computational scientists who do not have biological training. Appendix B provides information about the gene expression datasets that are used as examples in this book. Learning about analysis of DNA microarray data is facilitated by experience analyzing real data. Therefore, on our Web site `http://linus.nci.nih.gov/~brb` we provide the datasets used as examples in this book. Individuals can practice analyzing these datasets, or their own data, using the software of their choice. Appendix C describes the BRB-ArrayTools software. This software includes many of the methods described in the text and is regularly being extended with more tools based upon our experience in the analysis of microarray data. Again, BRB-ArrayTools is available on our Web site without charge for noncommercial purposes. It can

be licensed from the National Institutes of Health by commercial organizations.

2

DNA Microarray Technology

2.1 Overview

DNA microarrays are assays for quantifying the types and amounts of mRNA transcripts present in a collection of cells. The number of mRNA molecules derived from transcription of a given gene is an approximate estimate of the level of expression of that gene; see Appendix A for basic information on the biology of gene expression. RNA is extracted from the specimen and the mRNA is isolated. The mRNA transcripts are then converted to a form of labeled polynucleotides, called targets, and placed on the microarray. Details of the labeling process are provided later in this chapter.

The microarray consists of a solid surface on which strands of polynucleotides have been attached in specified positions. We refer to the polynucleotides immobilized on the solid surface as *probes*. The probes consist either of cDNA printed on the surface or shorter *oligonucleotides* synthesized or deposited on the surface. The labeled targets bind by hybridization to the probes on the array with which they share sufficient sequence complementarity. After allowing sufficient time for the hybridization reaction, the excess sample is washed off the solid surface. At that point, each probe on the microarray should be bound to a quantity of labeled target that is proportional to the level of expression of the gene represented by that probe. By measuring the intensity of label bound to each probe, one obtains numbers that, after adjustment for technical artifacts, should provide an estimate of the level of expression of all the corresponding genes.

2.2 Measuring Label Intensity

The amount of labeled target bound to each polynucleotide probe is quantified by illuminating the solid surface with laser light of a frequency tuned to the fluorescent label employed, and then measuring the intensity of fluorescence

over each probe on the array. This intensity of fluorescence should be proportional to the number of molecules of target bound to the probe. For a given number of bound molecules, other factors that can influence the intensity of fluorescence include the labeling efficiency, the number of polynucleotide strands in the probe, the laser voltage, and the photomultiplier tube setting. The number of bound molecules will be affected by the number of cells in the specimen, the RNA extraction efficiency, and the spatial distribution of labeled sample on the array.

The fluorescence emitted by molecules of targets bound to a probe is measured by a detector. Most commercial scanners use confocal microscopy detection. A confocal microscope focuses the photons originating in a very small region on the array to a photomultiplier tube. By collecting photons from one very small region at a time, the confocal method is effective in limiting contamination of the signal by other sources of fluorescence. This is important because the fluorescent signal emitted by the fluorophore is relatively weak. The resolution of most commercial confocal microscope-based scanners is about 3 μm, much less than the diameter of the region containing the probe. The array is scanned, collecting photons from each 3 μm region (pixel). At each step of the scan, the photons are focused into a photomultiplier tube where the photon density is translated into an electrical current which is amplified and digitized. If there are two samples cohybridized to the array with two fluorophores, the array is scanned for each label. With many systems, the array is scanned twice for each label and the average intensities recorded.

The fluorescent microscope does not directly measure the intensity of fluorescence over each probe. The instrument does not even know where the probes are located on the surface of the array. Instead, the microscope measures the intensity of fluorescence at each location of an imaginary grid covering the array surface. The grid locations are called *pixels*, short for picture elements. The distance between pixels is much less than the distance between probes. The output of the fluorescent microscope is a computer file, called an *image file*, giving the intensity of fluorescence measurement at each pixel. If two labeled samples were cohybridized, then two files are output, one corresponding to each label, or each *channel*. An image analysis algorithm processes these image files to estimate the intensity of label in each channel over each probe on the array, as described in Chapter 4.

2.3 Labeling Methods

For glass slide arrays the mRNA is usually reverse-transcribed to complementary DNA (cDNA), and a fluorescent label is incorporated into the cDNA during or after the reverse transcription reaction. The labeled cDNA is then placed on the microarray.

For Affymetrix GeneChip^{TM} arrays the preparation of labeled targets is somewhat different (Affymetrix 2000). After isolation of mRNA, cDNA is

synthesized. The cDNA is used as a template for T7 RNA polymerase to amplify the cDNA into synthesized cRNA molecules. In this amplification step a biotin label is introduced into the cRNA. The cRNA molecules are then fragmented into molecules 80 to100 nucleotides long. The biotin-labeled cRNA fragments are then hybridized to the GeneChip^{TM}. After hybridization, the bound cRNA fragments are stained with a biotin antibody.

2.4 Printed Microarrays

Microarrays differ in many important details. cDNA microarrays usually consist of probes of cDNA robotically printed on a microscope slide coated with poly-lysine or poly-amine to enhance absorption of the DNA probes (Schena et al. 1995, 2000). The robotic printers have several pins arranged in a rectangular pattern (Figure 2.1). For example, if there are four pins, then for

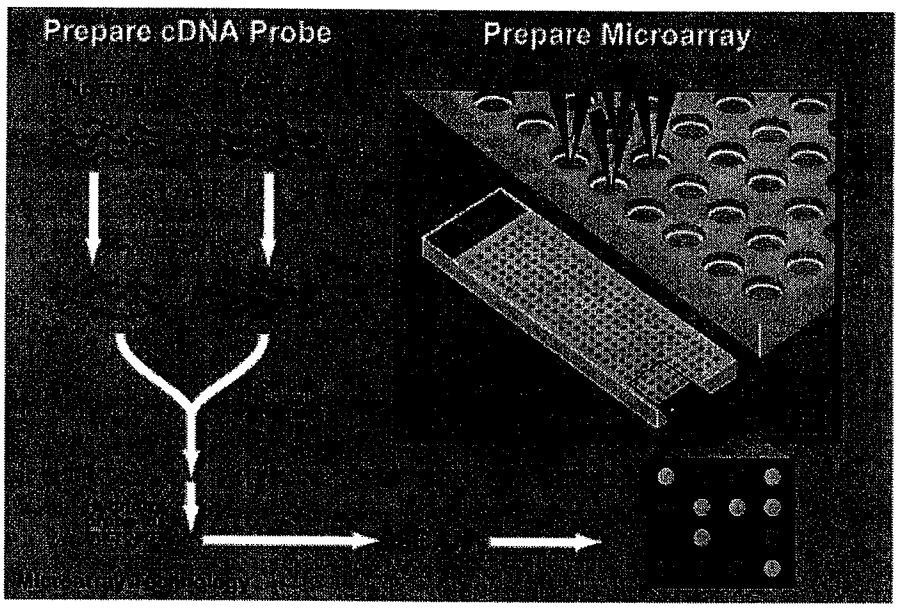

Fig. 2.1. Schematic of robotic printing of spots for cDNA array (right) and of processing of RNA samples for cohybridization to array.

each location of the robotic arm, four spots will be printed. At any time, the pins are loaded with cDNA from four different inventory wells and these PCR product clones are printed on each array of the print run. Then the pins are automatically washed and loaded with four other clones. The arm advances either horizontally or vertically an amount equal to the distance between spots, and the four clones are printed on all of the arrays of the print run. Thus,

for a four-pin printer, the spots on the array are printed in four rectangular grids corresponding to the rectangular arrangement of the robotic pins (Figure 2.2). The spots of each grid are printed with the same pin of the robot. The distance between the spots corresponds to the distance that the robotic arm moves between loadings of the pins, and the distance between the grids corresponds to the distance between the pins.

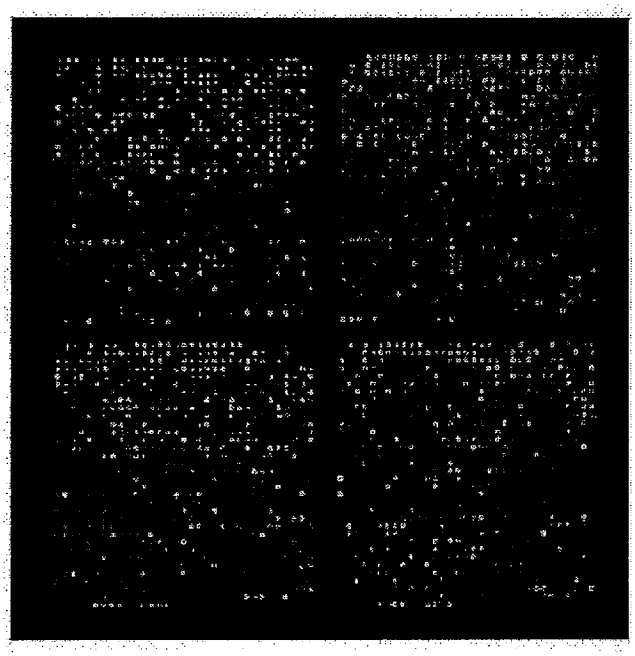

Fig. 2.2. A typical cDNA microarray image with two rows and two columns of grids.

Because the cDNA probes are generally several hundred bases long, stringent hybridization conditions can be employed and cross-reactivity is limited. However, robotic printing often results in substantial variability in the size and shape of corresponding spots on different arrays. Also, with cDNA arrays, the labeled sample is not uniformly distributed across the face of the array and the distribution of the sample differs among otherwise identical arrays. Hence direct comparison of intensities of corresponding probes on different arrays is problematic. Some of the interarray variability can be eliminated by a statistical "normalization" described in Chapter 4. Even after normalization, however, there is often substantial among corresponding variability spots on different arrays. Much of this variability can be controlled by co-hybridizing two samples on the same array. The two cDNA samples are labeled with differ-

ent fluorescent dyes. By using two laser sources, the intensity of fluorescence in each of the two frequency channels is measured over each probe. The second sample may represent either a specimen whose expression profile relative to the first specimen is of biological interest, or a reference sample used on all arrays in order to control experimental variability.

Externally synthesized oligonucleotides can also be robotically printed on coated glass slides. Because sample distribution across the face of the arrays remains variable, much of the interslide variability that is characteristic of cDNA arrays also applies to printed oligonucleotide arrays. Consequently, cohybridization of two separately labeled samples is also advantageous.

2.5 Affymetrix GeneChip™ Arrays

Affymetrix GeneChip™ arrays have oligonucleotide probes lithographically synthesized directly on the array. The array in this case is not a glass slide, but a silicon chip (Fodor et al. 1991). The oligonucleotides at all locations on the chip are synthesized in parallel. At the first step, the chip is bathed in a solution containing a precursor to one of the four nucleotides, say G. The synthesis of a nucleotide and attachment of the nucleotide to the anchor or the partially constructed oligonucleotide chain is light actuated. A mask is employed to ensure that light reaches only those addresses where the next nucleotide in the desired sequence is that represented by the current bath, say G. The in situ synthesis continues in this manner with multiple baths, washes, and masks employed.

The probes on GeneChip™ arrays are more homogeneous and less variable relative to cDNA arrays. Inter array variability due to sample distribution effects is also minimized because the samples are circulated inside the GeneChip™ during hybridization. Because of these reductions in inter array variability, a single sample is usually hybridized to GeneChips™.

The expense of fabrication and frequency of sequence errors for GeneChips™ increase with the length of the oligonucleotide probes employed, therefore relatively short 25 mer oligonucleotides are generally used. In order to obtain sufficient binding strength from 25 mer oligonucleotides, the hybridization conditions must be made less stringent than for cDNA arrays or longer spotted oligonucleotide arrays. Consequently, substantial cross hybridization is possible.

Affymetrix attempts to deal with the cross-hybridization problem by using multiple *probe pairs* for each target transcript (Lockhart et al. 1996). A probe pair consists of a 25 mer oligonucleotide perfectly complementary to a 25 nucleotide sequence of an exon of the target gene, and a 25 mer that differs from that perfect match probe by a single mismatched nucleotide at the central position. Affymetrix expects that the mismatched probe should not hybridize well to the target transcript but should hybridize to many transcripts to which the perfect-match oligonucleotide cross-hybridizes. Thus the intensity of signal

at the perfect match probe minus the intensity at the mismatched paired probe may be a better estimate of the intensity due to hybridization to the true target transcript.

Current GeneChips^{TM} use 11 to16 probe pairs for each target gene but the lengths of the probes are smaller than for cDNA arrays. The differences in perfect-match minus mismatch intensities are averaged across the probe pairs to give an estimate of intensity of hybridization to the target transcript; see Section 4.3.

2.6 Other Microarray Platforms

Several companies such as Protogene (Menlo Park, CA) and Agilent Technologies (Palo Alto, CA) in collaboration with Rosetta Inpharmatics (Kirkland, WA) have developed methods of in situ synthesis of oligonucleotides on glass arrays using ink-jet technology that does not require photolithography. The ink-jet technology of Agilent can also be used to attach pre synthesized DNA probes to glass slides.

Another class of DNA microarrays utilizes cDNA probes printed on a nylon membrane, and radioactive labeling of the sample. The radioactive label provides a stronger signal than fluorescent dye. This is useful when the amount of mRNA available for labeling is limited, but the wide scattering of label limits the density of probes that can be printed on the array, and larger format arrays are necessary. Although most of the principles of experimental design and analysis apply equally to arrays using radioactively labeled samples as to arrays using fluorescent labels, we generally talk in terms of the latter.

3

Design of DNA Microarray Experiments

3.1 Introduction

Microarray based experiments, like all experiments, should be carefully planned. Careful planning begins with a clear objective. The objective drives the selection of specimens and the specification of an appropriate analysis strategy. It is a common misconception that microarray experiments do not require planning or objectives; in this view, expression profiles are placed in a pattern recognition blackbox and discoveries emerge. Although pattern recognition algorithms have a role for some objectives involving microarrays, most successful microarray-based experiments have a definite focus.

There is substantial confusion about the role of "hypothesis testing" in studies using microarrays. It is true that microarray-based research is generally not based on a mechanistic biological hypothesis focused on specific genes. Other technologies are more suitable for testing hypotheses about specific genes. Nevertheless, most good microarray experiments are based on a hypothesis. For example, the hypothesis might be that there are genes whose expression is up-regulated or down-regulated in a tumor compared to normal tissue of the same tissue type. Or, the hypothesis might be that different tumors of the same tissue type and the same stage are not homogeneous with regard to gene expression profiles. Clearly identifying the general hypothesis of the study is important for ensuring that the type and number of specimens collected are appropriate. Clarity on the general hypotheses is also important for selecting methods of data analysis. A DNA microarray is just a highly parallel assay. It does not herald an era in which good practices of carefully thinking about the objectives of the experiment and of carefully planning the experiment and its analysis are obsolete.

Because DNA microarray investigations are not focused on a prespecified gene-specific hypothesis, there is much more opportunity for spurious findings than with more traditional types of investigations. Although the contexts in which microarrays are used are exploratory, strong claims are often made about which genes are differentially expressed under specified condi-

tions, which are disregulated in diseased tissue, and which are predictive of response to treatment. The serious multiplicity problems inherent in examining expression profiles of tens of thousands of genes mandate careful planning and special forms of analysis in order to avoid being swamped by spurious associations.

Design issues can be divided into those relating to the design of the DNA microarray assay itself and issues involving the selection, labeling, and arraying of the specimens to be assayed. In this chapter, we focus on the latter issues. Section 3.2 describes the importance of defining the study objectives for designing a microarray study. Section 3.3 discusses the difficulties in satisfying study objectives when only two RNA samples are compared. The sources of variation and the levels of replication of the experiment, discussed in Section 3.4, are important to consider when designing a study. Section 3.5 discusses the possibility of pooling samples and assaying the pooled sample with a microarray. With dual-label microarrays, the different ways of pairing and labeling the samples are discussed in Sections 3.6 and 3.7, respectively. The chapter ends with a discussion of the sample sizes required to meet the study objectives.

3.2 Study Objectives

DNA microarrays are useful in a wide variety of investigations with a wide variety of objectives. Many of these objectives fall into the following categories.

3.2.1 Class Comparison

Class comparison focuses on determining whether gene expression profiles differ among samples selected from predefined classes and identifying which genes are differentially expressed among the classes. For example, the classes may represent different tissue types, the same tissue under different experimental conditions, or the same tissue type for different classes of individuals. In cancer studies, the classes often represent distinct categories of tumors differing with regard to stage, primary site, genetic mutations present, or with regard to response to therapy; the specimens may represent tissue taken before or after treatment or experimental intervention. There are many study objectives that can be identified as class comparison. The defining characteristic of class comparison is that the classes are predefined independently of the expression profiles. Many studies are performed to compare gene expression for several types of class definition. For example, two genotypes of mice may be studied under two different experimental conditions. One analysis may address differences in gene expression for the two types of animals under the same experimental condition and the other analysis may address the effect of the experimental intervention on gene expression for a given genotype.

3.2.2 Class Prediction

Class prediction is similar to class comparison except that the emphasis is on developing a statistical model that can predict to which class a new specimen belongs based on its expression profile. This usually requires identifying which genes are informative for distinguishing the predefined classes, using these genes to develop a statistical prediction model, and estimating the accuracy of the predictor. Class prediction is important for medical problems of diagnostic classification, prognostic prediction, and treatment selection.

3.2.3 Class Discovery

Another type of microarray study involves the identification of novel subtypes of specimens within a population. This objective is based on the idea that important biological differences among specimens that are clinically and morphologically similar may be discernible at the molecular level. For example, many microarray studies in cancer have the objective of developing a taxonomy of cancers that originate in a given organ site in order to identify subclasses of tumors that are biologically homogeneous and whose expression profiles either reflect different cells of origin or other differences in disease pathogenesis (Alizadeh et al. 2000; Bittner et al. 2000). These studies may uncover biological features of the disease that pave the way for development of improved treatments by identification of molecular targets for therapy.

3.2.4 Pathway Analysis

The objective of some studies is the identification of genes that are coregulated or which occur in the same biochemical pathway. One widely noted example is the identification of cell cycle genes in yeast (Spellman et al. 1998). Pathway analysis is often based on performing an experimental intervention and comparing expression profiles of specimens collected before and at various time intervals after the experimental intervention. In some cases, however, pathway analysis may involve comparing the wild type organism to genetically altered variants.

3.3 Comparing Two RNA Samples

The initial cDNA microarray studies involved the cohybridization of one mRNA sample labeled with one fluorescent dye and a second mRNA sample labeled with a second fluorescent dye on a single microarray (DeRisi et al. 1996). This type of study, and the high cost of microarrays, left many investigators hoping and believing that no replication was needed. It also led to the publication of a variety of statistical methods for comparing the expression levels in the two channels at each gene on a single microarray. Even today,