Analyzing Microarray Gene Expression Data

This Page Intentionally Left Blank

Analyzing Microarray Gene Expression Data

WILEY SERIES IN PROBABILITY AND STATISTICS

Established by WALTER A. SHEWHART and SAMUEL S. WILKS

A complete list of the titles in this series appears at the end of this volume.

Analyzing Microarray Gene Expression Data

Geoffrey J. McLachlan
The University of Queensland
Department of Mathematics and
Institute for Molecular Bioscience
St. Lucia, Brisbane
Queensland, Australia

Kim-Anh Do
University of Texas
M. D. Anderson Cancer Center
Department of Biostatistics and
Applied Mathematics
Houston, Texas

Christophe Ambroise
U.M.R. C.N.R.S. Heudiasyc
Université de Technologie
de Compiègne
Compiègne, France

WILEY-
INTERSCIENCE

A JOHN WILEY & SONS, INC., PUBLICATION

Published by John Wiley & Sons, Inc., Hoboken, New Jersey.
Published simultaneously in Canada.

Limit of Liability/Disclaimer of Warranty: While the publisher and author have used their best efforts in preparing this book, they make no representation or warranties with respect to the accuracy or completeness of the contents of this book and specifically disclaim any implied warranties of merchantability or fitness for a particular purpose. No warranty may be created or extended by sales representatives or written sales materials. The advice and strategies contained herein may not be suitable for your situation. You should consult with a professional where appropriate. Neither the publisher nor author shall be liable for any loss of profit or any other commercial damages, including but not limited to special, incidental, consequential, or other damages.

For general information on our other products and services please contact our Customer Care Department within the U.S. at 877-762-2974, outside the U.S. at 317-572-3993 or fax 317-572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print, however, may not be available in electronic format.

Library of Congress Cataloging-in-Publication Data Is Available

ISBN 0-471-22616-5

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To

Beryl, Jonathan, and Robbie

Brad and Alex

Martine, Manon, Lison, and Liou

This Page Intentionally Left Blank

Contents

Preface

In recent times, there has been an explosion in the development of comprehensive, high-throughput methods for molecular biology experimentation. An example is the advent in DNA microarray technologies, such as cDNA arrays and oligonucleotide arrays, that provide means for measuring tens of thousands of genes simultaneously. These technologies benefit biological research greatly and further our understanding of biological processes by drawing together researchers in biology and quantitative fields including statistics, mathematics, computer science, and physics. In addition to the enormous scientific potential of microarrays to help in understanding gene regulation and interactions, microarrays have very important applications in pharmaceutical and clinical research.

This book has been written with two types of readers in mind: biologists who will undertake the statistical analyses of their own experimental microarray data, and biostatisticians entering the field of microarray gene expression data analysis. The primary focus of the book is on data analysis methods for this field; however, the biology and technology behind gene expression microarray experiments, as well as cleaning and normalization of the data, will be briefly covered.

Although biological experiments vary considerably in their design, the data generated by microarray experiments can be viewed as a matrix of expression levels, organized by genes versus tissue samples. In the case where a tissue sample corresponds to a single microarray experiments, we can represent the output from M experiments in the form of a $N \times M$ array (matrix). Each column of the matrix (the expression signature vector) contains the expression levels on the N genes monitored in the microarrays, while each row (the expression profile) contains the expression levels of a gene as it varies over the M tissue samples. Outside this matrix of expression levels, we may have covariate information for samples, genes, or both. The goal of microarray data analysis is to make inferences among samples, genes, and their expression levels and covariates.

The actual measurement of the expression levels raises several statistical issues in experimental design, image processing, outlier detection, transformations, and nonlinear modeling. We consider some of these issues (which are still ongoing as we complete this book) in the first two chapters. The rest of the book then considers the analysis of the microarray data, assuming that they have been appropriately preprocessed.

This analysis is centered on methods for the detection of differential expression, for cluster analysis (unsupervised classification), and for discriminant analysis (supervised classification) of microarray data.

An important and common question in microarray experiments is the detection of genes that are differentially expressed in tissue samples across a number of specified classes. These classes may correspond to tissues (cells) that are at different stages in some process, in distinct pathological states, or under different experimental conditions. A plethora of methods to detect differential gene expression are presented.

Cluster analyses have demonstrated their utility in the elucidation of unknown gene function, the validation of gene discoveries, and the interpretation of biological processes. Discriminant analysis is playing an ever-increasing role in predicting gene function classes and cancer classification.

There are two distinct clustering problems with microarray data. One problem concerns the clustering of the tissues on the basis of the genes. The clusters of tissues can play a useful role in the discovery and understanding of new subclasses of diseases. The second problem concerns the clustering of the genes on the basis of the tissues. The clusters of genes obtained can be used to search for genetic pathways or groups of genes that might be regulated together. Also, in the first problem above, we may wish first to summarize the information in the very large number of genes by clustering them into groups, which can be represented by some metagenes. We can then carry out the clustering of the tissues in terms of these metagenes.

In both the clustering of the tissues and the genes, hierarchical (agglomerative) clustering has been the most widely used method for the analysis of patterns of gene expression. It produces a representation of the data with the shape of a binary tree, in which the most similar patterns are clustered in a hierarchy of nested subsets. Nevertheless, classical hierarchical clustering presents drawbacks when dealing with data containing a non-negligible amount of noise, as is the present case. Also, there is no reason why the clusters of tissues or genes should belong to a hierarchy such as in the evolution of species. In this book, the emphasis is on a model-based approach to clustering. An advantage of model-based clustering is that it provides a sound mathematical framework for clustering. In particular, it provides a principled statistical approach to the practical questions that arise in applying clustering methods, namely, the question of what metric (distance function) to adopt and the question of how many clusters there are in the data.

In recent times, model-based clustering has become very popular in the statistical literature. Unfortunately, as the data to be analyzed from microarray experiments often have gene-to-sample ratios of approximately 100-fold, off-the-shelf parametric methodology does not apply at least to the classification of the tissues on the basis of the genes. This is because the dimension of the feature space (the number of genes) is so much greater than the number of observations (the number of tissues). But even the cluster analysis of the genes on the basis of the tissues is a nonstandard problem, as the genes are not all independently distributed.

An obvious way to handle the very large number of genes is to perform a principal component analysis (PCA) and carry out the cluster analysis on the basis of the leading components. But a potential problem with a PCA is the determination of an appropriate number of principal components (PCs) useful for clustering. A common practice is to choose the first few leading components. But it is not clear where to stop and whether some of these components are caused by some artifact or noises in the data unrelated to the clustering task. Also, there is the difficulty of interpretation of components because each component has loadings generally on all genes.

Hence the focus in the book is on the EMMIX-GENE procedure, which is a normal mixture-based method of clustering that has been especially developed for the clustering of tissue samples or other high-dimensional data. This procedure has an option for an initial selection of the genes where genes that appear to have little clustering capacity are discarded. It then clusters the (standardized) gene profiles into groups, effectively using Euclidean distance as the metric, with the aim that highly correlated genes are put in the same cluster. Each group of genes is then represented by a single metagene (the group-sample mean) and then clustering is performed in terms of the metagenes. This divide-and-conquer approach is becoming popular in the bioinformatics literature for both unsupervised and supervised classification of tissue samples.

The clustering step of EMMIX-GENE makes use of, if needed, mixtures of factor analyzers. That is, it provides a global nonlinear approach to dimension reduction as it postulates a finite mixture of linear submodels (factor models) for the distribution of the full signature vector or a reduced version of metagenes given the (unobservable) factors. Thus, it is a local dimensionality reduction method in contrast with a PCA, which is a global linear method.

A number of discriminant rules are discussed for the supervised classification of the tissue samples. However, the aim of this book is not to provide a comprehensive review of available methods but rather to focus on what we think are useful methods for the analysis of microarray data. To this end, the focus in discriminant analysis of tissue samples is on the support vector machine. It has the advantage that it can be formed from all the genes and its performance is generally not too disadvantaged as a consequence of using all the genes. Its performance can be improved by undertaking feature selection using an easily implemented procedure called recursive feature elimination.

In the statistical analyses, including discriminant and cluster analyses, some form of feature selection will usually be carried out. A consequence of basing the final analysis on a selected "top" subset of the available genes is that there will typically be a selection bias that needs to be corrected for in relating the conclusions to subsequent (new) data. In the case of a discriminant rule, it means that the selection bias has to be allowed for in the estimation of the generalization error. Otherwise, a false overoptimistic impression will be obtained for the discriminatory power of the rule.

This bias has often been overlooked in the bioinformatics literature. Also, this bias arises in an unsupervised context with tests and plots on the number of clusters.

The first two chapters of this book aim to (1) provide a bridge to the biological and technical aspects involved in microarray experiments, and (2) summarize and emphasize the need for basic research in DNA array technologies and statistical thinking through every step of the microarray experiment and analysis to enhance reliability and reproducibility of research results.

Chapter 1 is an introductory chapter and provides a review on DNA microarrays and relevant technology. In particular, we begin with the biological principles behind microarray experiments. Background information on the substrates and technology used in microarray gene expression studies is intended for the biostatistician who is not familiar with the biological experiments. We discuss DNA, cDNA, oligonucleotides, and the development of microarray technology, as well as the steps involved in the manufacture of cDNA microarrays and in generating experimental microarray data. Commercial arrays, primarily the GeneChip®, are also briefly introduced in this chapter.

Chapter 2 discusses cleaning and normalization of gene expression microarray data, as well as the need for designs of experiments with replicated data.

Chapter 3 considers in a general context some methods for the cluster analysis of multivariate data consisting of n independent observations taken on a p-dimensional feature vector associated with the random phenomenon of interest. The focus is on model-based methods of clustering and it covers the use of mixtures of factor analyzers for high-dimensional data such as microarray data. In relation to the problem of how many clusters there are in the data, consideration is given to the problem of assessing the number of components in a mixture model by resampling.

Chapter 4 considers the development of the model-based methodology covered in Chapter 3 for its application to problems in the clustering of tissue samples. The emphasis is on the EMMIX-GENE procedure which has been developed specifically for the clustering of tissue samples. Its application to real microarray data sets is illustrated on two well-known sets in the literature. Also, it is demonstrated on several real data sets how this model-based approach to clustering can be used to consider the question of how many clusters of tissues there are in the data.

Chapter 5 focuses on the selection of differentially expressed genes in known classes of tissue samples. As this problem concerns the selection of significant genes from a large pool of candidate genes, it needs to be carried out within the framework of multiple hypothesis testing. The recent and fruitful literature on the latter topic in the context of microarrays is covered in depth. Distributional problems, including use of the t-distribution and its variants to provide robustness are introduced with a discussion of numerous methods, frequentist and Bayesian, to handle the multiplicity issue. The latter part of this chapter considers the clustering of genes that have been identified as being differentially expressed with a view to finding: (1) groups of genes that are significantly correlated with each other; (2) groups of genes that share similar expressions across the tissues.

Chapter 6 considers methods in discriminant analysis or supervised classification in a general context with a view to their application to microarray data. Discriminant rules covered include the traditional normal-based linear and quadratic discriminant classifiers, more flexible parametric rules based on normal mixtures or mixtures of factor analyzers, support vector machines and their variants, nearest-neighbor and nearest centroid rules, classification trees, and neural networks. The problem of error-rate estimation of a discriminant rule is considered too, along with ways for the provision of standard errors for the estimates of the error rates.

Chapter 7 considers applications of some of the discriminant rules introduced in the previous chapter to the supervised classification of tissue samples. In applications concerned with the diagnosis of cancer, one class may correspond to cancer and the other to benign tumors. In applications concerned with patient survival following treatment for cancer, one class may correspond to the good prognosis group and the other to the poor prognosis group. Also, there is interest in the identification of "marker" genes that characterize the different tissue classes. Attention is focused on applications of the support vector machine and nearest-shrunken centroids, which is a recent version of nearest centroids to handle the very large number of genes. These two approaches are demonstrated on some cancer data sets. Particular attention is paid to the need to correct for the selection bias in estimating the prediction capacity of a discriminant rule formed from a subset of genes selected from a much larger set.

Chapter 8 is concerned with linking results of a model-based clustering of tumor tissues on cancer biology and clinical outcome. Cancer patients with the same stage of disease can have markedly different treatment responses and clinical outcome. Thus there is much interest in whether microarray expression data can be used to provide prognostic information beyond that provided by stage and other traditional clinical criteria. We report some recent results that show that the clustering provides significant prognostic information on the outcome of the disease beyond that available in current systems based on histopathology criteria and extent of disease at presentation.

The authors wish to acknowledge several Houston-based colleagues. LeeAnn Chastain contributed significant efforts to Chapter 1 with respect to the literature search, writing, editing, and proofreading. We are grateful to Wei Zhang, Keith Baggerly, Kevin Coombes, Li Zhang, and Tuyet-Trinh Do for reading and commenting of this chapter. Peter Müller was a great collaborator on the nonparametric Bayesian mixture model for differential gene expression. Sijin Wen assisted the second author in programming the gene shaving method. Bradley Broom contributed significantly with advanced programming using his high-performance computing tool.

Concerning acknowledgments to colleagues in Brisbane, the authors wish to thank Nazim Khan, Abdollah Khodkar, Katrina Monico, and Justin Zhu for their assistance. They wish to acknowledge the significant contributions made by Angus Ng and Liat Ben-Tovim Jones to the research leading to the results reported in Chapter 8. Further thanks are due to Liat for her many very helpful comments and suggestions on drafts of the manuscript. Finally, special thanks are due to Richard Bean who has greatly assisted with all aspects of the book, including proof reading and overseeing the numerous technical issues that arose during the preparation of the manuscript in camera-ready form.

The first author was supported by the Australian Research Council. The second author was partially supported by University of Texas SPORE in Prostate Cancer grant

CA90270, and the Early Detection Research Network grant CA99007. Thanks are due too to the authors and owners of copyright material for permission to reproduce tables and figures, and to Joel Tyndall (Institute for Molecular Bioscience) who used the InsightII Modeling Environment (Accelrys, 2004) to prepare the art work for the cover of the book.

Brisbane, Australia	Geoff McLachlan
Houston, USA	Kim-Anh Do
Compiègne, France	Christophe Ambroise

1

Microarrays in Gene Expression Studies

1.1 INTRODUCTION

Recently, the scientific world has witnessed an explosion in the development of comprehensive, high-throughput methods for molecular biology experimentation. Potentially, these cutting-edge techniques will allow researchers to characterize genetic diseases such as cancer at the molecular level, and will lead to new treatments directed at specific cellular aberrations. The focus in this book is on the output from array technologies, which have made it straightforward to monitor simultaneously the expression pattern of thousands of genes. We are concerned with how to analyze such massive data sets.

In this chapter, we provide background information on the substrates and technology used in microarray gene expression studies. It is intended for biostatisticians who are not familiar with the biological experiments that produce their microarray data. We discuss DNA, cDNA, oligonucleotides, and the development of microarray technology as well as the steps involved in the manufacture of cDNA microarrays and in generating experimental microarray data. Commercial arrays, primarily the GeneChip®, are also introduced briefly in this chapter. In Chapter 2 we discuss cleaning and normalization of gene expression microarray data and their effects on methods for detecting differential expression. Subsequent chapters of the book are devoted to statistical analyses of the data taken to be cleaned and normalized.

1.2 BACKGROUND BIOLOGY

1.2.1 Genome, Genotype, and Gene Expression

The human genome is a representation of our entire gene complement. The human genome map, completed in April 2003, represents the identification and prediction of the base-pair sequences along each of the 23 pairs of chromosomes present in the human cell nucleus. Even as researchers celebrated the completion of the map ahead of its scheduled date, the human genome was not (and is not) a known entity. In addition to chromosomal areas that still prove difficult to map, a multitude of unknown variations in the genome complicate the identification of individual gene complements. In fact, scientists use a different term when speaking about one person's complete gene complement: *genotype*. Each person's genotype may be unique because there are untold numbers of genetic sequence variations in the form of mutations and polymorphisms.

A related research endeavor, the International HapMap project, started in October 2002, will identify and describe the patterns of variations in DNA sequences that are common among humans. This research involves identifying the sites in the human genome where persons differ by a single base (known as a single nucleotide polymorphism, or SNP), and identifying sets of associated SNPs, known as haplotypes. The ultimate goal of this project is to produce a database of the common haplotypes in the human genome and the SNPs that can be used as tags for each of the haplotypes. (See `http://hapmap.org` for additional information.)

The initial mapping of the human genome has provided a common foundation to which researchers in the field of genetics and in the many overlapping fields of molecular biology, biochemistry and biophysics, biostatistics, pharmacogenetics, bioinformatics, computer science, and many others will contribute from this point forward.

The development of computational models and methods for the investigation of gene expression patterns has already led to important biostatistical research projects, and its importance will continue to grow because of the increasing specialization of biomedicine. The greatest biomedical gains are being realized through knowledge of a specific subtype of a disease or disorder, the specific biochemical pathways affected by the disease and by the therapy prescribed, and the myriad characteristics of the individual patient's genotype and phenotype[1] that result in his or her very unique biological response to the disease or disorder and to the therapy that is prescribed.

What is to be gained from the measurement of gene expression patterns? Experiments are designed to observe the changes in a gene in response to external stimuli and/or to the activation or expression of other genes, allowing the observation and measurement of the relative expression of a gene. Cell samples are exposed experimentally to human hormones, toxins, pharmacologic agents, and so on, and the resulting increase or decrease in the transcription (expression) of a particular DNA

[1] A phenotype comprises all the physical, biochemical, and physiological characteristics of a person as determined through genetic and environmental influences. A phenotype is also the manifestation or expression of a gene or gene pair in human characteristics.

segment or gene can be measured. This information will be used to elucidate the potential pathways of genes as well as the interrelations among various genes. It will be applied to the development of pharmacologic agents and genetic therapies with a level of target specificity that is well beyond that of our current ability to analyze disorders, implement preventive measures, or prescribe medical treatments appropriately. Scientists are learning that complex disorders result from the interactions of many genes and are identifying the components of the interactions.

1.2.2 Of Wild-Types and Other Alleles

Researchers in the human genome project have determined that chromosome 20 is made up of approximately 60 million bases and contains 727 genes (Hattori and Taylor, 2001). It is believed that a human being inherits 30,000 to 40,000 genes from each parent. What is a gene? A *gene* is a specific segment of a DNA molecule that contains all the coding information necessary to instruct a cell to synthesize a specific product, such as an RNA molecule or a protein. Contained within the gene are segments that we acknowledge as active in the coding process *(exons)*, as well as segments that are noncoding *(introns)*. Each gene also represents a basic unit of a person's biological inheritance from his or her two parents. Genes can be "mapped" because each occupies a specific location (or locus) on a chromosome, and each chromosome can be specifically identified as well.

Genes are identified according to their apparent general or specific function. It is believed that *housekeeping genes* (for example, GAPDH, B-actin, tubulin) are expressed or functional in all cells because they encode proteins that are needed for basic cellular activity. Additional examples of gene types that have been identified include the immunoglobulin genes, which code to direct the synthesis of specific types of immunoglobulins (antibodies); and a tumor suppressor gene (antioncogene), which functions to limit the formation and growth of malignant cells. By definition, human genes function to promote and regulate biological activity that is considered necessary and productive for the functioning of the organism. It is not correct to state that a gene codes for a disease or predisposes a person to a specific disorder. Rather, it is a deleterious mutation in a gene that may predispose a person to a specific disease or disorder.

A variation or any alternative form of a gene that is found to occupy the same locus on a particular chromosome is known as an *allele*. A *wild-type allele* is the form of a particular gene that is thought to have developed through the evolutionary processes that exist in nature (called "wild" because it is a product of nature itself). A gene that is found to have a mutation will be labeled as a specific allele of that gene, which is different from the wild-type allele and from other alleles that identify other types of mutations occurring at that same chromosomal locus.

1.2.3 Aspects of Underlying Biology and Physiochemistry

Deoxyribonucleic acid (DNA) is contained within chromosomes in the nucleus of each cell. The DNA molecule consists of two anti-parallel strands of sugar–phosphate linkages that are bonded together in a right-handed double helix by the noncovalent hydrogen bonding between pairs of attached amino bases, which lie in a flat plane roughly perpendicular to the long axis of the molecule. The anti-parallel arrangement of the nucleotide chains requires the transcription of a new RNA or DNA chain to run in the opposite direction of the template. Hydrophobic interactions between the stacked bases in the interior of the DNA molecule also stabilize the double helix by packing it tightly to exclude water and other nonpolar molecules. Adenine, thymine, guanine, and cytosine are the amine bases, the sequential order of which contributes to the functioning of a particular segment of the DNA strand (a gene). The bases exhibit a characteristic and specific bonding known as *base pairing*. Base pairing (also known as *Watson–Crick base pairing*) is a chemical bonding process that allows molecular hybridization to occur. Between two strands of DNA, the base known as adenine (A) specifically bonds with thymine (T) through two hydrogen bonds, and guanine (G) specifically bonds to cytosine (C) through two hydrogen bonds, in a manner that creates the double helix. Between a strand of DNA and a strand of ribonucleic acid or RNA (during transcription), adenine from the DNA strand will bond specifically to the base uracil (U) from the RNA strand, and guanine will again bond specifically to cytosine. The amine base that will form a bonding pair with another amine base (A with T or A with U, and G with C) is considered to be its complementary base, and a single strand of DNA or RNA that contains the same sequential order of complementary bases for bonding as a given strand is considered to be its complementary strand. Single DNA or RNA strands will form stable bonds only with a complementary strand. This specificity of bonding allows the "message" of the sequence of base pairing in that segment of DNA to be communicated through the process of transcription.

Transcription is the communication of a genetic code from DNA to RNA through the synthesis of a strand of RNA that has sequences of bases complementary to that of the DNA strand. Genetic transcription is carried out to direct the activity of the cell. The sequence of the bases in a DNA segment comprises the code or genetic instructions that are passed on from the DNA molecule to the RNA molecule because of the specific pairing that occurs between the bases in DNA and RNA. Nucleic acids that guide the production of proteins [2] are transcribed in the nucleus of the cell as messenger RNA (mRNA). Microarray technology utilizes these properties of specific bonding or *hybridization* of a single strand of DNA to a complementary strand of DNA or RNA. The hydrogen bonding between the bases is relatively weak and can be broken by heating the DNA or RNA sample to its melting temperature (approximately 90 °C) through a process referred to as *denaturing*. The single denatured strands of the polynucleotide can then be attached to a solid substrate or used to probe strands

[2] The process of synthesizing polypeptide chains from mRNA is known as *translation*, wherein the sequence of bases in the mRNA strand determines the amino acid sequence in the protein that is produced.

of unknown coding order in experiments. Once the denatured DNA is slowly cooled to approximately 60°C, *reassociation* occurs. Reassociation is the process whereby single strands of the polynucleotide associate with complementary strands through random collisions, resulting in the formation of specific amine base pairs through hydrogen bonding. Reassociation is facilitated if the DNA sample is fragmented into short lengths of nucleotides, thus increasing the number of random collisions and increasing the probability that complementary chains will undergo base pairing.

1.3 POLYMERASE CHAIN REACTION

Polymerase chain reaction (PCR) is a technique that "amplifies" or replicates DNA fragments. It is commonly used to create billions of copies of specific fragments of DNA from a single DNA molecule. This technique has numerous applications in medical research, in forensic science, and in many related fields and is used to produce DNA for the manufacture of microarrays. The PCR technique was developed in 1983 through the work of Kary B. Mullis, a biochemist, and his colleagues at Cetus Corporation in Emeryville, California (Mullis, 1990). [F. Hoffman–LaRoche Ltd. and Roche Molecular Systems, Inc., purchased the patent for the PCR technique from Cetus Corporation; however, its recognition as an acceptable patent, since it is based on a naturally occurring enzyme, is currently under dispute in United States appeals courts. European courts upheld the patent in a ruling issued in 2003. See Dalton (2001) and Knight (2003).]

The PCR technique is based on the catalytic action of a DNA polymerase enzyme that is stable at high temperatures, such as those used to denature DNA and RNA molecules. The initial technique utilized a DNA polymerase enzyme isolated from the genetically engineered bacterium *Thermus aquaticus (Taq)*, which was found in thermal springs of Yellowstone National Park in Wyoming. Use of a polymerase enzyme from bacteria with characteristics similar to the *Taq* bacteria enables the DNA replications to be conducted at high temperatures for fast reaction rates and can be rigorously controlled for high fidelity. In human cell division, a primer (a short RNA segment that functions to start the copying of the DNA strands) starts the creation of a template of each single strand of DNA in each chromosome as the base pairing bonds separate. The polymerase then takes over, creating the DNA templates that reproduce the genetic material in the creation of a new cell. For the PCR technique, a *Taq* polymerase from the bacterium is provided, along with the primers and a supply of the four nucleotide bases (adenine, guanine, cytosine, and thymine). The DNA to be duplicated is then added to a vial containing these components. The vial is heated to 90°C for 30 seconds to denature the DNA, separating the strands. The vial is then slowly cooled to 60°C to allow the primers to bind to the DNA strands, and it is again heated to promote the action of the *Taq* polymerase. The entire process, duplicating each piece of DNA in the vial, takes less than 2 minutes. The cycle is then repeated for the same vial approximately 30 times, with each new DNA segment acting as a new template, exponentially reproducing the number of DNA segments in the vial (Mullis, 1990). Recombinant *Taq* polymerase, obtained by the insertion of the gene for the *Taq* polymerase into another type of bacteria (and currently held under a second patent by F. Hoffman–LaRoche Ltd.), is now more commonly used for DNA amplification (Dalton, 2001).

Following the PCR process, the DNA samples are purified to reduce the presence of unwanted components as well as salts and primers used in the PCR process. Purification is done by precipitation, gel-filtration chromatography, or both (Duggan et al., 1999). PCR products representing specific genes are then applied to the array to manufacture DNA microarrays.

1.4 CDNA

Messenger RNA (mRNA) is the form of ribonucleic acid that directs the production of cellular proteins, so it is important in experiments of gene expression. Researchers want to observe what cellular proteins are produced and the function of those proteins in particular types of cells (such as tumor cells) or in response to specific external stimuli, so they are interested in testing the expression patterns of the mRNA. Although protein synthesis and activation are not regulated solely at mRNA levels in a cell, mRNA measurement is used to estimate cellular changes in response to external signals or environmental changes. The mRNA in a biological sample is first chemically bound to a DNA molecule in order to remove it from the other cellular components. The molecule of mRNA is relatively fragile, however, and can easily be broken down by the action of enzymes that are prevalent in biological solutions, so researchers commonly manipulate a form of DNA that possesses the complementary bases of the mRNA while existing in a more stable state. This form of DNA, known as *complementary DNA* (cDNA), is created directly from the sample mRNA through a procedure known as *reverse transcription* (transcribing complementary genetic base sequences from RNA to DNA). cDNA is also called *synthetic DNA*, since it is formed through reverse transcription from RNA rather than through self-replication during cell division. cDNA is generally prepared in strand lengths of 500 to 5,000 bases of known sequence.

1.4.1 Expressed Sequence Tag

Human genes contain base-pairing sequences that are replicated, as well as sequences that are not replicated, during mRNA translation to form specific polypeptide chains in protein synthesis. The sequences that are translated in protein synthesis are coding sequences, known as *exons*, while the noncoding sequences are known as *introns*. Enzymes activated during mRNA transcription recognize the noncoding junctions in the nucleotide sequence and splice together the exons for protein production after removing the introns. *Expressed sequence tag* (EST) is the name given to a short sequential segment from a gene. It is generated to represent the coding portion of a gene; thus, an EST is frequently used as a gene substitute for PCR amplification, microarray production, and experiments. Substituting shorter nucleotide sequences for genomic DNA was proposed in the 1980s and was first undertaken in experiments on cDNA clones derived from human brain tissue by a research group at the National Institute of Neurological Disorders and Stroke, National Institutes of Health in the United States (Adams and Bischof, 1994). ESTs are generated through transcription cloning from both ends of a cDNA sequence, through what is called incomplete *unedited single-pass sequencing reads* of cDNA, resulting in frequent errors (Marra et al., 1998). EST data can be used in general evaluations of gene expression but are not considered suitable for gene expression studies that require greater detail. ESTs have been shown to be valuable in facilitating gene identification and in genome mapping, and EST data comprise the bulk of most public DNA sequence databases (Gerhold and Caskey, 1996; Marra et al. 1998; Quackenbush, 2001; Wolfsberg and Landsman, 2001). The criticism of ESTs in gene libraries has been due primarily to an overabundant representation in the data of genes that are frequently expressed, resulting in redundancies, and an absence of representation of genes that are rarely expressed. Researchers generally try to correct for the presence of redundant EST data in a gene library.

1.5 MICROARRAY TECHNOLOGY AND APPLICATION

High-density DNA microarray technology allows researchers to monitor the interactions among thousands of gene transcripts in an organism on a single experimental medium, which is often a glass microscope slide or nylon membrane. Prior to the computerization and miniaturization of this technology, researchers were limited to examinations of much smaller numbers of genetic units per experiment and were able to assess interactions among genes under changing conditions on a much smaller scale. Microarray technology is particularly useful in the evaluation of gene expression patterns in complex disorders because of its ability to observe the expression of the same genes in different samples at the same time and in response to the same stimuli.

The use of microarrays in biomedical research is equivalent to some of the technological advancements found in the computer science industry, such as that of parallel distribution. Distributing the "work" of an experiment in a parallel fashion facilitates solving computationally complex problems and becomes more than the equivalent of running thousands of experimental steps at the same time. Microarrays are generally designed to provide parallel distribution of the work of an experiment. Each microarray can represent thousands of separate biochemical assays performed in a much shorter time period.

Microarrays can be used to evaluate the dynamic expression of genes in response to normal cellular activity (for example, changes in gene transcription, cell division) or in response to external stimuli (for example, a toxic substance, viral infection). The ability to simulate a large variety of cellular conditions and then translate and process the resulting large quantities of data, provides a systematic way to evaluate cellular function and genetic variations and may be particularly important in testing for genetic susceptibility to diseases and disorders as well as genetic susceptibility (or the ability to respond effectively) to specific therapies or interventions.

The biostatistician's concern lies in the statistical methods and computations that are required to appropriately normalize, analyze, and interpret the vast amounts of data obtained from gene expression studies using microarrays. It is important, however, for the biostatistician to develop a basic understanding of the procedures involved in production of the arrays and in the experiments that generate gene expression data. An understanding of how the data will be applied in a biomedical context is also an important factor. A biostatistician's initial task is to consider the appropriate statistical normalization [3] procedures that may need to be performed on data that are generated from microarray experiments. Understanding the components of the microarray experiments and the levels of sample processing are the crucial preliminary requirement that will guide the biostatistician (Nguyen et al., 2002; Kerr, 2003; Simon et al., 2002; Dobbin et al., 2003). Chapter 2 focuses on data normalization techniques and relevant controversies. The present chapter will provide the biostatistician with some basic principles underlying the microarray technology, beginning with a review of some terms and concepts common to studies that use DNA microarrays.

1.5.1 History of Microarray Development

Microarray technology developed through the application of advanced technologies from the fields of biology and physiochemistry to the analysis of ligand assays, particularly those involving immunoassays. Assays are determinations of the amount of a particular substance within a mixture of different substances. For example, assays have been in use for decades to identify blood proteins; to test for chemical exposure; to perform urinalyses; to screen for drugs; to screen for certain congenital mutations (such as α-fetoprotein); to test for blood clotting disorders; to measure antibody titers; and to test for enzymes specific to injury to the heart muscle or liver tissue. Immunoassays help determine the amount of antibodies present in a biological sample that are involved in the very specific antibody–antigen binding that occurs in immunologic response processes. Researchers in this field were among the first to introduce microarray technology. Labeling techniques implemented in immunoassays included fluorescent labeling of either the antibody or the antigen to detect its presence, as well as radioactive labeling and enzyme-linked immunosorbent assay (ELISA).

Immunoassay technology, as developed in the 1950s and 1960s, involved the attachment of antibodies to solid supports and relied on the specificity of target molecules binding to the antibody (Polsky-Cynkin et al. 1985; Ekins, 1998). These same techniques would subsequently be adapted for DNA analysis. Early assays utilized macroarray technology, whereby the samples were applied or "spotted" manually onto a test surface, creating sample spot sizes of 300 μm or more. Once arrays were designed to support "sample spots" of less than 200 μm in diameter; however,

[3] Normalization is the process of standardizing the data so that reasonable data comparisons can be made.