Exploration and Analysis of DNA Microarray and Other High-Dimensional Data
Exploration and Analysis of DNA Microarray and Other High-Dimensional Data

Second Edition

DHAMMIKA AMARATUNGA
Janssen Pharmaceutical Companies of Johnson and Johnson
USA

JAVIER CABRERA
Rutgers University
USA

ZIV SHKEDY
Hasselt University
Belgium

WILEY
To our families in America, Sri Lanka, Spain, and Israel
Contents

Preface xv

1 A Brief Introduction 1
  1.1 A Note on Exploratory Data Analysis, 3
  1.2 Computing Considerations and Software, 4
  1.3 A Brief Outline of the Book, 5
  1.4 Data Sets and Case Studies, 6
    1.4.1 The Golub Data, 6
    1.4.2 The Mouse5 Data, 7
    1.4.3 The Khan Data, 7
    1.4.4 The Sialin Data, 7
    1.4.5 The Behavioral Study Data, 8
    1.4.6 The Spiked-In Data, 8
    1.4.7 The APOAI Study, 8
    1.4.8 The Breast Cancer Data, 8
    1.4.9 Platinum Spike Data Set, 9
    1.4.10 Human Epidermal Squamous Carcinoma Cell Line A431 Experiment, 9
    1.4.11 Note: Public Repositories of Microarray Data, 9

2 Genomics Basics 10
  2.1 Genes, 10
  2.2 Deoxyribonucleic Acid, 11
  2.3 Gene Expression, 12
  2.4 Hybridization Assays and Other Laboratory Techniques, 14
2.5 The Human Genome, 16
2.6 Genome Variations and Their Consequences, 17
2.7 Genomics, 19
2.8 The Role of Genomics in Pharmaceutical Research and Clinical Practice, 19
2.9 Proteins, 22
2.10 Bioinformatics, 23
Supplementary Reading, 24

3 Microarrays
3.1 Types of Microarray Experiments, 26
  3.1.1 Experiment Type 1: Tissue-Specific Gene Expression, 26
  3.1.2 Experiment Type 2: Developmental Genetics, 26
  3.1.3 Experiment Type 3: Genetic Diseases, 27
  3.1.4 Experiment Type 4: Complex Diseases, 28
  3.1.5 Experiment Type 5: Pharmacological Agents, 28
  3.1.6 Experiment Type 6: Plant Breeding, 29
  3.1.7 Experiment Type 7: Environmental Monitoring, 29
3.2 A Very Simple Hypothetical Microarray Experiment, 30
3.3 A Typical Microarray Experiment, 31
  3.3.1 Microarray Preparation, 32
  3.3.2 Sample Preparation, 33
  3.3.3 The Hybridization Step, 33
  3.3.4 Scanning the Microarray, 34
  3.3.5 Interpreting the Scanned Image, 34
3.4 Multichannel cDNA Microarrays, 35
3.5 Oligonucleotide Microarrays, 36
3.6 Bead-Based Arrays, 37
3.7 Confirmation of Microarray Results, 37
Supplementary Reading and Electronic References, 37

4 Processing the Scanned Image
4.1 Converting the Scanned Image to the Spotted Image, 39
  4.1.1 Gridding, 40
  4.1.2 Segmentation, 40
  4.1.3 Quantification, 41
4.2 Quality Assessment, 42
  4.2.1 Visualizing the Spotted Image, 43
## CONTENTS

4.2.2 Numerical Evaluation of Array Quality, 44  
4.2.3 Spatial Problems, 45  
4.2.4 Spatial Randomness, 46  
4.2.5 Quality Control of Arrays, 47  
4.2.6 Assessment of Spot Quality, 48  

4.3 Adjusting for Background, 49  
4.3.1 Estimating the Background, 49  
4.3.2 Adjusting for the Estimated Background, 52  

4.4 Expression-Level Calculation for Two-Channel cDNA Microarrays, 53  

4.5 Expression-Level Calculation for Oligonucleotide Microarrays, 53  
4.5.1 The Average Difference, 54  
4.5.2 A Weighted Average Difference, 54  
4.5.3 Perfect Matches Only, 55  
4.5.4 Background Adjustment Approach, 56  
4.5.5 Model-Based Approach, 56  
4.5.6 Absent–Present Calls, 57  

Supplementary Reading, 58  
Software Notes, 58  

5 Preprocessing Microarray Data  
5.1 Logarithmic Transformation, 60  
5.2 Variance Stabilizing Transformations, 62  
5.3 Sources of Bias, 63  
5.4 Normalization, 63  
5.5 Intensity-Dependent Normalization, 65  
5.5.1 Smooth Function Normalization, 68  
5.5.2 Quantile Normalization, 68  
5.5.3 Stagewise Normalization, 71  
5.5.4 Normalization of Two-Channel Arrays, 71  
5.5.5 Spatial Normalization, 76  
5.5.6 Normalization of Oligonucleotide Arrays, 76  
5.6 Judging the Success of a Normalization, 77  
5.7 Outlier Identification, 80  
5.8 Nonresistant Rules for Outlier Identification, 80  
5.9 Resistant Rules for Outlier Identification, 81  
5.10 Assessing Replicate Array Quality, 83  
Software Notes, 83
# 6 Summarization

6.1 Replication, 86  
6.2 Technical Replicates, 87  
6.3 Biological Replicates, 91  
6.4 Experiments with Both Technical and Biological Replicates, 91  
6.5 Multiple Oligonucleotide Arrays, 95  
6.6 Estimating Fold Change in Two-Channel Experiments, 96  
6.7 Bayes Estimation of Fold Change, 97  
6.8 Estimating Fold Change Affymetrix Data, 98  
6.9 RMA Summarization of Multiple Oligonucleotide Arrays Revisited, 100  
6.10 Factor Analysis for Robust Microarray Summarization, FARMS, 105  

## Software Notes 106

# 7 Two-Group Comparative Experiments

7.1 Basics of Statistical Hypothesis Testing, 111  
7.2 Fold Changes, 112  
7.3 The Two-Sample \( t \)-Test, 114  
7.4 Diagnostic Checks, 116  
7.5 Robust \( t \)-Tests, 118  
7.6 The Mann–Whitney–Wilcoxon Rank Sum Test, 119  
7.7 Multiplicity Adjustment: The Familywise Error Rate, 121  
7.7.1 A Pragmatic Approach to the Issue of Multiplicity, 122  
7.7.2 Simple Multiplicity Adjustments, 123  
7.7.3 Sequential Multiplicity Adjustments, 123  
7.8 Multiplicity Adjustment: The False Discovery Rate, 125  
7.8.1 Benjamini and Hochberg Procedure, 125  
7.8.2 The Positive False Discovery Rate, 129  
7.9 Resampling-Based Multiple Testing Procedures, 130  
7.10 Small-Variance-Adjusted \( t \)-Tests and SAM, 132  
7.10.1 Modifying the \( t \)-Statistic, 135  
7.10.2 Assessing Significance with the SAM \( t \) Statistic, 135  
7.10.3 Strategies for Using SAM, 138  
7.10.4 An Empirical Bayes Framework, 139  
7.10.5 Understanding the SAM Adjustment, 139  
7.11 Conditional \( t \), 141  
7.12 Borrowing Strength Across Genes, 145  
7.12.1 Simple Methods, 147
CONTENTS

7.12.2 A Bayesian Model, 148
7.13 Two-Channel Experiments, 149
7.13.1 The Paired Sample t Test and SAM, 150
7.13.2 Borrowing Strength Via Hierarchical Modeling, 150
7.14 Filtering, 152
7.14.1 Filtering Based on Summarized Data, 152
7.14.2 Filtering Based on Probe-Level Data, 156
Supplementary Reading, 160
Software Notes, 161

8 Model-Based Inference and Experimental Design Considerations 166
8.1 The F-Test, 167
8.2 The Basic Linear Model, 168
8.3 Fitting the Model in Two Stages, 171
8.4 Multichannel Experiments, 171
8.5 Experimental Design Considerations, 172
8.5.1 Comparing Two Varieties with Two-Channel Microarrays, 172
8.5.2 Comparing Multiple Varieties with Two-Channel Microarrays, 174
8.5.3 Single-Channel Microarray Experiments, 176
8.6 Miscellaneous Issues, 177
8.7 Model-Based Analysis of Affymetrix Arrays, 177
8.7.1 One-Way ANOVA, 177
8.7.2 Linear Models for Microarray Data (Limma), 179
8.7.3 A Joint Model for Gene Expression and Response, 184
8.7.4 Analysis of Dose–Response Microarray Experiments, 187
8.7.5 Analysis of Time Course Data, 191
Supplementary Reading, 195
Software Notes, 196

9 Analysis of Gene Sets 200
9.1 Methods for Identifying Enriched Gene Sets, 201
9.1.1 MLP and Fisher’s Test, 202
9.1.2 GSEA and the Kolmogorov–Smirnov Test, 205
9.2 ORA and Fisher’S Exact Test, 205
9.3 Interpretation of Results, 206
9.4 Example, 206
Software Notes, 206
10 Pattern Discovery

10.1 Initial Considerations, 210
10.2 Cluster Analysis, 212
   10.2.1 Dissimilarity Measures and Similarity Measures, 213
   10.2.2 Guilt by Association, 215
   10.2.3 Hierarchical Clustering, 216
   10.2.4 Partitioning Methods, 222
   10.2.5 Model-Based Clustering, 226
   10.2.6 Chinese Restaurant Clustering, 227
   10.2.7 Ensemble Methods for Clustering Samples, 228
   10.2.8 Discussion, 229
10.3 Seeking Patterns Visually, 229
   10.3.1 Principal Components Analysis, 230
   10.3.2 Factor Analysis, 234
   10.3.3 Biplots, 237
   10.3.4 Spectral Map Analysis, 237
   10.3.5 Multidimensional Scaling, 239
   10.3.6 Projection Pursuit, 239
   10.3.7 Data Visualization with the Grand Tour and Projection Pursuit, 241
10.4 Biclustering, 242
   10.4.1 Block Clustering, 243
   10.4.2 Gene Shaving, 244
   10.4.3 The Plaid Model, 244
Supplementary Reading, 247
Software Notes, 247

11 Class Prediction

11.1 Initial Considerations, 251
   11.1.1 Misclassification Rates, 251
   11.1.2 Reducing the Number of Classifiers, 252
11.2 Linear Discriminant Analysis, 258
11.3 Extensions of Fisher’s LDA, 261
11.4 Penalized Methods, 263
11.5 Nearest Neighbors, 264
11.6 Recursive Partitioning, 265
   11.6.1 Classification Trees, 267
   11.6.2 Activity Region Finding, 269
CONTENTS

11.7 Ensemble Methods, 270
  11.7.1 Random Forest, 271
  11.7.2 Enriched Random Forest, 272
11.8 Enriched Ensemble Classifiers, 273
11.9 Neural Networks, 273
11.10 Support Vector Machines, 276
11.11 Generalized Enriched Methods, 277
  11.11.1 Enriched Principal Components Analysis and Biplots, 279
  11.11.2 Enriched Penalized Methods: Lasso, SVM, P-SVM, 281
  11.11.3 Enriched Partial Least Squares (PLS), 283
11.12 Integration of Genome Information, 285
  11.12.1 Integration of Gene Expression Data and Molecular Structure Data, 285
  11.12.2 Pathway Inference, 285
Supplementary Reading, 286
Software Notes, 286

12 Protein Arrays 290

12.1 Introduction, 290
12.2 Protein Array Experiments, 291
12.3 Special Issues with Protein Arrays, 292
12.4 Analysis, 293
12.5 Using Antibody Antigen Arrays to Measure Protein Concentrations, 294

References 298

Index 313
Preface

In August 1999, at the Joint Statistical Meetings in Baltimore, two of us [DA & JC] were invited to present a paper on “The Analysis of DNA Microchip Data.” This was the first presentation in any Joint Statistical Meetings on the topic of DNA microarrays, as they are now called. In just a few years, the field has exploded and, in each of the recent Joint Statistical Meetings, there have been over a hundred presentations related to DNA microarrays!

The Baltimore paper outlined many of the issues that are still being discussed today, including intensity-dependent normalization, the use of methods that are robust to outliers, improving the sensitivity of the analysis by borrowing strength across genes, and interpretation of results in a way that is germane to the overall objectives of a study. However, there have been many developments since then, and this book represents our effort at organizing this material into a semicoherent whole.

ACKNOWLEDGMENTS

We would like to thank several bioinformaticians, scientists, and statisticians at Johnson and Johnson and elsewhere who patiently helped to educate us about genomics and microarrays and in writing this book. We would particularly like to acknowledge Jim Colaianne, Luc Bijnens, Hinrich Gohlmann, Peter Ghazal, Michael McMillian, Nandini Raghavan, Zhenya Cherkas, Albert Lo, Kwok Tsui, David Tyler, Birol Emir, Alicia Nieto Reyes, Chin-Ray Yu, Yung-Seop Lee, Volha Tryputsen, Willem Talloen, An de Bondt, Hugo Ceulemans, Tomasz Burzykowski, Tim Davison, Jackson Wan, Xuejun Liu, Harindra Abeyesinghe, Gordon Pledger, Sepp Hochreiter, Adetayo Kasim, Martin Otava, Willem Talloen, Suzy Van Sanden, Dan Lin, Nolen Joy Perualila, Tatsiana Khamikova, Pushpika Thilakaratne, and Dani Yekutieli. We would also like to note that numerous graduate students at Rutgers University and Universiteit Hasselt assisted us in our research.
Ziv Shkedy gratefully acknowledge support from the Belgian IUAP/PAI network “Statistical techniques and modeling for complex substantive questions with complex data.” And last, but not least, we would like to thank Gayatri Amaratunga and Maria Drelich, who helped in various ways.

Dhammika Amaratunga
Javier Cabrera
Ziv Shkedy
CHAPTER 1

A Brief Introduction

Data analysis has, quite suddenly, begun to assume a prominent role in the life sciences. From being a science that generally produced relatively limited amounts of quantitative data, biology has, in the space of just a few years, become a science that routinely generates enormous amounts of it.

To a large part, this metamorphosis can be attributed to two complementary advances. The first is the successful culmination of the Human Genome Project and other genome sequencing efforts, which have generated a treasure trove of information about the DNA sequences of the human genome and the genomes of several other species, large and small. This has resulted in a huge number of genes being newly identified and biologists are now confronted with the daunting, but exhilarating, task of ascertaining their functions.

This is where the second advance, the emergence of modern experimental technology, such as microarray technology, comes in. Currently, the most widely used form of this technology is the DNA microarray, which offers scientists the ability to monitor the behavior patterns of several thousands of genes simultaneously, allowing them to study how these genes function and follow how they act under different conditions. Another form of microarray technology, the protein array, provides scientists the capability of monitoring thousands of proteins simultaneously, for similar purposes. And this is just the beginning. Emerging technical innovations, such as bead-based arrays, have the potential to increase throughput much more.

These developments have ushered in a thrilling new era of molecular biology. Traditional molecular biology research followed a “one gene per experiment” paradigm. This tedious and inherently exhausting approach was capable of producing only limited results in any reasonable period of time. Although it has, without question, logged a series of remarkable achievements over the years, this approach does not allow anything close to a complete picture of gene function and overall genome behavior to be readily determined.
The advent of microarray technology has created an opportunity for doing exactly this by fast-tracking research practice away from a “one gene” mode to a “thousands of genes per experiment” mode and allowing scientists to study how genes function, not just each on their own, but jointly as well.

In fact, the way microarray technology is revolutionizing the biological sciences has been likened to the way microprocessors transformed the computer sciences toward the latter part of the twentieth century (through miniaturization, integration, parallel processing, increased throughput, portability, and automation) and the way the computer sciences, in turn, transformed many other disciplines just a few years later. Microarray technology has been brought into play to characterize genomic function in genome systems spanning all the way from yeast to human.

Microarray experiments are conducted in such a manner as to profile the behavior patterns of thousands of nucleic acid sequences or protein simultaneously. Plus, they are capable of being automated and run in a high-throughput mode. Thus, they can, and do, generate mountains of data at an ever increasing pace. Thus, the proper storage, analysis, and interpretation of this data have turned out to be a major challenge.

Our focus is on the analysis part. After all, data by itself does not constitute knowledge. It must be first be analyzed and relationships and associations studied and confirmed, in order to convert it into knowledge. By doing so, it is hoped that a complete picture of the intermeshing patterns of biomolecular activity that underlie complex biological processes, such as the growth and development of an organism and the etiology of a disease, would emerge.

One issue is that the structure of the data is singular enough to warrant special attention. The raw data from a DNA microarray experiment, for example, is a series of scanned images of microarrays that have been subjected to an experimental process. The general plan for analyzing this data involves converting these images into quantitative data, then preprocessing the data to transform it into a format suitable for analysis, and, finally, applying appropriate data analysis techniques to extract information pertinent to the biological question under study. Application of statistical methodology is feasible as these experiments can be run on replicate samples, although, by and large, the amount of replication tends to be limited. Thus, a complexity is that, while there is data on thousands and thousands of genes, the information content per gene is small. As a result, there is a sense that much of the data collected in microarray experiments remains to be fully and properly interpreted.

It should, therefore, not be a surprise that statistical and computational approaches are beginning to assume a position of greater prominence within the molecular biology community. While these quantitative disciplines have a rich and impressive array of tools to cover a very broad range of topics in data analysis, the structure of the data generated by microarrays is sufficiently unique that, either standard methods have to be tailored for use with microarray data, or an entirely fresh set of tools have to be developed specifically to handle such data. What has happened, of course, is a confluence of the two. The purpose of this book is to present an extensive, but, by no means, exhaustive, series of
computational, visual, and statistical tools that are being used for exploring and analyzing microarray data.

1.1 A NOTE ON EXPLORATORY DATA ANALYSIS

Early statistical work was essentially enumerative and exploratory in nature. Statisticians were concerned with developing effective ways of discerning patterns in quantitative data. Then, from about a fourth of the way into the twentieth century, mathematics-driven confirmatory techniques began to dominate the field of statistics, driving data exploration into the background. The focus began to be the development of optimal ways to analyze data rigorously, but under various sets of fairly restrictive assumptions.

Fortunately, toward the latter part of the twentieth century, data exploration began to make a comeback as an imperative aspect of statistics, having been revitalized almost single-handedly by Tukey (1962, 1977, 1986), who likened it to detective work. Exploratory data analysis (EDA), as the modern incarnation of statistical data exploration is called, is an approach for data analysis that employs a range of techniques (many graphical), in a strategic manner, in order to

- gain insight into a data set;
- discover systematic structures, such as clusters, in the data;
- flag outliers and anomalies in the data;
- assess what assumptions about the data are reasonable.

The last of these guides the data analyst to an approach or a model that should be suitable for a more formal phase in the analysis of the data. This confirmatory data analysis (CDA) phase, which may involve inferential procedures such as confidence interval estimation and hypothesis testing, allows the data analyst to probabilistically model the uncertainties of a situation to assess the reproducibility of the findings. By doing so, CDA ensures that chance patterns are not mistaken for real structure. Even at this phase, however, EDA stresses the importance of running diagnostic checks to assess the validity of any underlying assumptions (e.g., Anscombe and Tukey, 1963; Daniel and Wood, 1971).

EDA is particularly well suited to situations where the data is not well understood and the problem is not well specified, such as screening. Because of this, EDA techniques have found their way into the world of data mining (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). In such situations, broad-based methods that have the ability to discover and illustrate essential aspects of the data are of most value. Proper data visualization tools, for instance, are highly effective both at revealing facets of the data that otherwise may not have been apparent and at challenging assumptions about the data that otherwise may have been taken for granted.

It could be argued that EDA is as much an attitude or a philosophy about how a data analysis should be conducted as an assortment of techniques. The EDA
A BRIEF INTRODUCTION

approach suggests strategies for carefully scrutinizing a data set: how to examine a
data set, what to look for, and how to interpret what has been observed. The key is
that EDA permits the data itself to reveal its underlying structure and model without
the data analyst having to make too many possibly indefensible assumptions.

Over the years, the popularity of EDA has been boosted by a number of notewor-
thy publications by Tukey and his students and colleagues, such as Mosteller and
Tukey (1977), Velleman and Hoaglin (1981), Hoaglin (1982), Hoaglin, Mosteller,
Fernholz, and Morgenthaler (1997), Fernholz, Morgenthaler and Stahel (2001),
and Cabrera and McDougall (2002), and has gained a large following as the most
effective way to seek structures in data. Hoaglin, Mosteller, and Tukey (1983) is
an excellent introduction to EDA.

That is not to forget CDA. Tukey (1980) argues that exploratory and confirmatory
analyzes must both be components of a good data analysis. This is the
approach we will take in this book.

1.2 COMPUTING CONSIDERATIONS AND SOFTWARE

The data analyst must have access to computing resources, both hardware and
software, that are capable of dealing with the huge amounts of data that must be
analyzed. Holloway et al. (2002) is a review of some of the issues related to this
topic.

A number of software packages offer the data analyst powerful tools for EDA
and CDA, including interactive graphics and a large collection of statistical pro-
cedures. Two that are commonly used in the analysis of microarray data are \texttt{R}
(Ihaka and Gentleman, 1996) and \texttt{SPLUS}. Other statistical packages that are good
for EDA include SAS, JMP, DataDesk, Matlab, and MINITAB.

In addition, libraries of routines specially designed for analysis of microarray
data have begun to spring up. Some of these are in the public domain, others are
only available commercially. A couple are listed below.

- **DNAMR** (URL: \url{http://www.rci.rutgers.edu/cabrera/DNAMR}),
  which stands for “DNA Microarray Routines,” is a collection of \texttt{R} programs
developed by the authors of this book and their collaborators. Implementations
  of many of the procedures described in this book are available in the DNAMR
  package and can be downloaded from the book’s web page.

- **Bioconductor** (URL: \url{http://www.bioconductor.org}) is a collection
  of software packages for the analysis and interpretation of DNA microarray
  and other high-dimensional data. It is free, open source, and open develop-
  ment. While it is also based primarily on \texttt{R}, it does contain contributions in
  other programming languages.

Care must be taken that software is not used blindly—using the wrong methods to
analyze data from an experiment could produce meaningless “findings” and miss
signals of potential interest. Unfortunately, few, if any, off-the-shelf packages offer a comprehensive data handling system that integrates all the data-related needs, such as data acquisition, storage, extraction, quality assurance, and analysis, that are essential for even a moderate-sized microarray laboratory.

1.3 A BRIEF OUTLINE OF THE BOOK

This book will mainly focus on the analysis of data from microarray experiments. However, the concepts underlying the methods presented here, if not the methods themselves, can be used in other settings that generate high-dimensional or megavarate data. Such settings are becoming more and more commonplace in biomedical research with the advent of new high-throughput technologies such as protein arrays, flow cytometry, and next-generation sequencing, in addition to microarrays. All these technologies generate large complex high-dimensional data sets from individual experiments. The complexity of this data highlights the importance of implementing proper data management and data analysis procedures as this will significantly impact the reproducibility or nonreproducibility of any findings from these experiments.

EDA and CDA techniques can be applied to microarray data (and other high-dimensional data) to

- assess the quality of a microarray;
- determine which genes are differentially expressed;
- classify genes based on how they coexpress;
- classify samples based on how genes coexpress.

Following this, the investigator will generally try to

- connect differentially expressed genes to annotation databases;
- locate differentially expressed genes on pathway diagrams;
- relate expression levels to other cell-related information;
- determine the roles of genes on the basis of patterns of coexpression.

Hopefully, this process will culminate in an insight of interest.

Figure 1.1 shows, schematically, the path of a typical microarray data analysis. Readers may find it useful to periodically refer to it. In this book, we will present a collection of techniques for analyzing microarray data as per the first set of bullets. Before we embark on our journey, a brief road map of where we are going may be helpful.

Chapter 2 is a brief introduction to molecular biology and genomics. Chapter 3 describes DNA microarrays, what they are, how they are used, and how a typical DNA microarray experiment is performed. Chapter 4 outlines how the output of a DNA microarray experiment, the scanned image, is processed and quantitated and
A BRIEF INTRODUCTION

Interpret and report findings
Analyze data
Check quality of normalized data
Normalize data
Transform data
Adjust for background
Check quality of spotted image
Analyze data
Interpret and report findings

Figure 1.1 Schematic of a typical microarray data analysis.

how image and spot quality checks are done. Chapter 5 discusses preprocessing microarray data, which typically involves transforming the data and then applying a normalization. Chapter 6 discusses summarization of data across replicates. Chapter 7 describes statistical methods used for analyzing the simplest comparative experiments, those involving just two groups. Chapter 8 discusses more complex experiments and issues related to their design. Chapter 9 discusses methods for analyzing gene sets. The next two chapters deal with multivariate methods: Chapter 10 discusses unsupervised classification methods and Chapter 11 discusses supervised classification methods. Chapter 12 describes protein arrays; a typical protein array experiment is outlined and methodology for analyzing protein array data is described.

1.4 DATA SETS AND CASE STUDIES

The following data sets are used at various points in this book.

1.4.1 The Golub Data

The Golub data first appeared in a seminal publication (Golub et al., 1999) which demonstrated that gene expression profiling could be used to classify malignancies. The study involved profiling the expression of several thousand genes in bone marrow from 38 patients with acute leukemia, consists of 27 with the acute lymphoblastic leukemia (ALL) and 11 with the acute myeloid leukemia (AML). In the
original analysis, 50 genes whose expression levels differed most between AML and ALL cells were found to be able to correctly classify which patients had AML and which had ALL in a blinded new cohort of 36 acute leukemia patients. The data set has gene expression measurements for 3051 genes for the original 38 tumor mRNA samples and is available in R.

1.4.2 The Mouse5 Data

The Mouse5 data are from an experiment involving 10 pairs of microarrays, C1A, C1B, C2A, C2B, ..., C10A, C10B. Each pair of microarrays corresponds to a single mRNA sample (labeled C1, C2, ..., and C10), which was taken from a mouse following treatment and hybridized to two separate microarrays (labeled A and B). The two microarrays in each pair are technical replicates as they were exposed to the same biological sample. The five mice from which samples C1–C5 were drawn are controls, so they are biological replicates, while each of the other five was treated with one of five drugs. There were 3300 genes arrayed on the microarrays.

1.4.3 The Khan Data

The Khan data set contains gene expression measurements, obtained using cDNA microarrays, from four types of pediatric small round blue cell tumors (SRBCTs): neuroblastoma (NB), rhabdomyosarcoma (RMS), the Ewing family of tumors (EWS), and Burkitt lymphomas (BL), a subtype of non-Hodgkin’s lymphoma. The four cancer types are clinically and histologically similar, yet their response to treatment is markedly different, making accurate diagnosis essential for proper therapy. The purpose of the study was to classify, as accurately as possible, a cell as being one of these four types using gene expression information. The microarrays measured the expression levels of 6567 genes. This data was filtered to remove any gene that consistently expressed below a certain minimum level of expression, leaving expression data for 2308 genes. A total of 88 cells were analyzed. Data for 63 of these cells (23 EWS, 20 RMS, 12 NB, 8 BL) were used as a training set, while the data for the remaining 25 cells (6 EWS, 5 RMS, 6 NB, 3 BL, 5 non-SRBCT) were set aside to make up a blind test set.

1.4.4 The Sialin Data

The Sialin data are from an experiment in which the gene expression profiles of mice whose Slc17A5 gene (which is responsible for the production of Sialin) was knocked out were compared to the gene expression profiles of wild-type mice (i.e., “normal” mice). In the experiment, RNA samples from total brain were derived from newborn and 18-day-old mice for each of the two groups: Slc17A5 knockout and wild type. There were six biological samples in each group for a total of 24 samples. Microarray experiments were performed on the RNA samples using Affymetrix Mouse430-2 GeneChips; the expression levels of 45,101 genes were profiled.
1.4.5 The Behavioral Study Data

This case study was obtained from a preclinical behavioral experiment, in which 24 male experimentally naive Long-Evans rats from Janvier (France) were randomized into two treatment groups with 12 rats per group. The first group of rats received placebo, while the second group was treated with a novel antipsychotic compound. Animals were tested in a large open field. Rat behavior data parameters were recorded systematically. In particular, the parameter of primary interest to the study investigators was the total distance traveled by the rats. Active response to the treatment was expected to increase this distance. In addition, microarray data, measuring the gene expression of 5644 genes, were obtained. Further details of the experiment can be found in Lin et al. (2012).

1.4.6 The Spiked-In Data

The Affymetrix HGU-133A Spiked-in data set is publicly available for the purpose of determining the sensitivity and specificity of various methods for the analysis of microarray data. The data set has an advantage over real-life data sets because the true number of differentially expressed genes is known. It contains known genes that are spiked-in at 14 different concentrations ranging from 0 to 512 pM, arranged in a Latin squared design. There are 42 arrays and 42 spiked-in probe sets equally distributed over the 14 concentrations. In addition to the original spiked-in transcripts, McGee and Chen (2006) discovered 22 additional probe sets that have similar characteristics as the spiked-in probe sets. Thus, the HGU-133A spiked-in data set contains 64 spiked-in probe sets out of the 22,300 probe sets.

1.4.7 The APOAI Study

This data set is from a cDNA experiment which compared the gene expression profiles of eight apolipoprotein AI (APOAI) knockout mice with the gene expression profiles of eight wild-type mice. Target mRNA was obtained from the liver tissue of each mouse and labeled using a Cy5 dye. The RNA from each mouse was hybridized to a separate microarray. Common reference RNA was labeled with Cy3 dye and used for all the arrays. This data has been employed by a number of authors to illustrate normalization methods for cDNA data. The data is available in R.

1.4.8 The Breast Cancer Data

A time course microarray data set. This data set comes from a breast cancer cell line microarray study. Data are available in six time points: 1, 4, 12, 24, 36, and 48 h after treatment. At each time point, eight replicate arrays are available, consisting of 1900 genes. For more details about the data, we refer to Lobenhofer et al. (2002) and Peddada et al. (2003).
1.4.9 Platinum Spike Data Set

The Platinum Spike data set consists of 18 Affymetrix GeneChip Drosophila Genome 2.0 microarrays, divided into two groups (A and B). The 2189 RNAs are spiked in so that the number of up- and downregulated transcripts between the two conditions is balanced. Moreover, 3426 transcripts were spiked in at the same concentrations for both conditions. The spiked-in transcripts are measured by 5615 probe sets, whereas the 13,337 probe sets are empty and not expected to carry any signal. In each group, there were three samples synthesized and for each sample, three technical replicates were used for the microarray analysis. For more details about the data, we refer to Zhu et al. (2010).

1.4.10 Human Epidermal Squamous Carcinoma Cell Line A431 Experiment

The data come from an oncology experiment designed to better understand the biological effects of growth factors in human tumor. Human epidermal squamous carcinoma cell line A431 was grown in Dulbecco’s modified Eagle’s medium, supplemented with L-glutamine (20 mM), Gentamicin (5 mg/ml,) and 10% fetal bovine serum. The cells were stimulated with growth factor EGF (R&D Systems, 236-EG) at different concentrations (0, 1, 10, and 100 ng/ml) for 24 h. RNA was harvested using RLT buffer (Qiagen). All microarray-related steps, including the amplification of total RNAs, labeling, hybridization, and scanning, were carried out as described in the GeneChip Expression Analysis Technical Manual, Rev.4 (Affymetrix, 2004). The collected data were quantile normalized in two steps: first within each sample group, and then across all sample groups obtained (Amaratunga and Cabrera, 2000, 2003; Bolstad et al., 2002). The resulting data set consists of 12 samples, 4 dose levels, and 3 microarrays at each dose level, with 16,998 probe sets.

1.4.11 Note: Public Repositories of Microarray Data

Bioconductor, ArrayExpress, and Gene Expression Omnibus (GEO) are large repositories of DNA microarray and other high-dimensional data.
CHAPTER 2

Genomics Basics

It is useful to review the basic concepts of modern molecular biology before fully immersing ourselves in the world of microarrays. We are sure that readers who have had limited exposure to this fast-developing field would appreciate this review, others may skip ahead. Genomics is a fascinating subject; after all, it is the story of life, and can occupy a multivolume book just by itself. In the interest of space, of course, it is necessary that we confine our discussion to those topics that are essential to an understanding of the science underlying microarrays, leaving other topics for interested readers to explore on their own. Some excellent general references that we, not being trained as molecular biologists ourselves, have found useful are listed at the end of the chapter.

2.1 GENES

From ancient times, it was suspected that there existed some sort of a hereditary mechanism that carried information from parent to child. It is because of this mechanism that family members tend to exhibit similar characteristics or traits. For example, they tend to resemble each other in terms of appearance and physical characteristics such as skin color; they tend to be predisposed toward certain diseases such as diabetes, cancer, and heart disease; and so on. However, inheritance is clearly not a perfect copying process. For example, a child of brown-eyed parents could turn out to be blue-eyed. Despite the efforts over the years of many leading scientists and thinkers to understand the hereditary mechanism, its precise nature remained an intriguing mystery until quite recently.

Following centuries of speculation and research, the existence of discrete hereditary units, which we now call genes, has been firmly established. Each gene, either by itself or in combination with some other genes, provides a clear and
unambiguous set of instructions for producing some property of its organism. The complete set of genes in an organism, essentially the master blueprint for that organism, is referred to as its genome. This blueprint contains all the hereditary instructions for building, operating, and maintaining the organism, and for passing life in like form on to the next generation of that organism.

Until the twentieth century, there was hardly any concrete information as to what genes were and how they operated. Then, a panoply of innovative research work and pathbreaking discoveries over (roughly) the first half of the twentieth century gave genes a chemical (molecular) existence. This culminated in the pivotal realization that genes are made of deoxyribonucleic acid (DNA).

2.2 DEOXYRIBONUCLEIC ACID

A DNA molecule consists of two long strands wound tightly around each other in a spiral structure known as a double helix. The structure has been likened to a twisted ladder, whose sides are made of sugar and phosphate and whose rungs are made of bases.

Each strand of the DNA molecule (i.e., each side of the ladder once it has been untwisted and straightened out) is a linear arrangement of repeating similar units called nucleotides. Every nucleotide has three components: a sugar (deoxyribose), a phosphate group, and a nitrogenous base. The base is one of adenine (A), thymine (T), guanine (G), or cytosine (C). The bases on one strand are paired with the bases on the other strand according to the complementary base pairing rules (also called the Watson–Crick base pairing rules): adenine only pairs with thymine, guanine only pairs with cytosine. The pairs so formed are called base pairs (bp); they form the coplanar rungs of the ladder. The force that holds a bp together is a weak hydrogen bond. Even though each individual bond is weak, their cumulative effect along the strands is strong enough to bind the two strands tightly together. As a result, DNA is chemically inert and is a stable carrier of genetic information.

The sequences of bases along each of the two strands of DNA are complementary to each other as they are matched by the complementary base pairing rules. This complementary sequencing has an important consequence. It was recognized from very early on that, whatever the entity was that was a hereditary unit, it must be able to self-replicate so that information could be passed on from generation to generation. At the time that the structure of DNA was deduced, there was a lot of excitement, as it was clear that the complementary structure of the DNA molecule would allow every DNA molecule to create an exact replica of itself, thus fulfilling this requirement.

The DNA replication process is, in principle, quite straightforward. First, the DNA molecule unwinds and the “ladder” unzips, thereby disrupting the weak bonds between the bps and allowing the strands to separate. Then, each strand directs the synthesis of a brand new complementary strand, with free nucleotides matching up with their complementary bases onto each separated strand, a process that produces
two descendant DNA molecules. Each descendant consists of one old and one new DNA strand. The constraints imposed by the complementary base pairing rules ensure that each new strand is an exact copy of the old one with the order of the bases along the strands being faithfully preserved.

The preservation of the base order is crucial. The particular order of the bases arranged along any one strand, its DNA sequence, is the mechanism that specifies the exact genetic instructions required to create the traits of a particular organism.

Many genes are located along each long DNA molecule. A gene is a specific contiguous subsequence of the DNA sequence whose A-T-G-C sequence is the code required for constructing a protein. Proteins are giant complex molecules made of chains of amino acids and it is they that are actually both the building blocks and the workhorses of life. Proteins also regulate most of life’s day-to-day functions; in fact, even the DNA replication process is mediated by enzymes, proteins whose job is to catalyze biochemical reactions.

2.3 GENE EXPRESSION

An organism’s DNA is located in its cells. Cells are the fundamental units of all living organisms, both structurally and functionally. A cell is a microscopic, yet extraordinarily complex, structure that contains a heterogeneous mix of substances essential to life.

There are many substructures within a cell. The most prominent one is a highly protected subcompartment called the nucleus, in which resides the organism’s DNA. Enclosing the nucleus is the nuclear membrane, the protective wall that separates the nucleus from the rest of the cell, which is called its cytoplasm. The entire cell is enclosed by the plasma membrane. Embedded within this membrane is a variety of protein structures that act as channels and pumps to control the movement of molecules into and out of the cell.

The set of protein-coding instructions in the DNA sequence of a gene resembles a computer program. A computer program must first be compiled and executed in order for anything to happen. In much the same way, a gene must be expressed in order for anything to happen. A gene expresses by transferring its coded information into proteins that dwell in the cytoplasm, a process called gene expression.

The transmission of genetic information from DNA to protein during gene expression is formulated by the central dogma of molecular biology, which can be stated in oversimplistic terms as “DNA → mRNA → protein.” This postulates that the protein-coding instructions from a gene are transmitted indirectly through messenger RNA (mRNA), a transient intermediary molecule that resembles a single strand of DNA. There are a few differences between mRNA and DNA, three being that mRNA is single stranded, its sugar is ribose, and it has the base uracil (U) rather than the base thymine.