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.

Exploration and Analysis of DNA Microarray and Other High-Dimensional Data, Second Edition.
Dhammika Amaratunga, Javier Cabrera, Ziv Shkedy.

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 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 noteworthy publications by Tukey and his students and colleagues, such as Mosteller and Tukey (1977), Velleman and Hoaglin (1981), Hoaglin (1982), Hoaglin, Mosteller, and Tukey (1983), Chambers et al (1983), Tukey (1986), Miller (1986), Brillinger, 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 procedures. Two that are commonly used in the analysis of microarray data are `R` (Ihaka and Gentleman, 1996) and `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: `http://www.rci.rutgers.edu/cabrera/DNAMR`), which stands for "DNA Microarray Routines," is a collection of `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: `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 development. While it is also based primarily on `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 megavariate 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

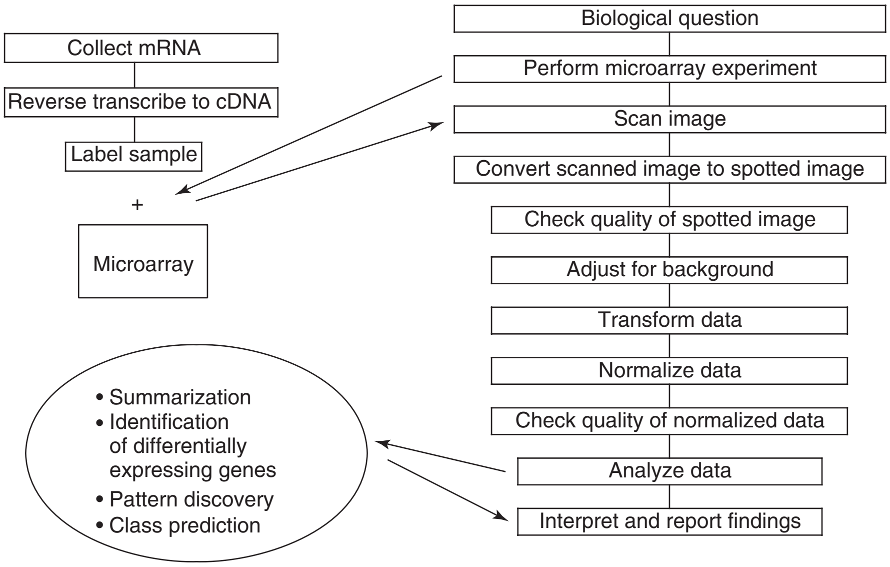
Biological question
Perform microarray experiment
Scan image
Convert scanned image to spotted image
Check quality of spotted image
Adjust for background
Transform data
Normalize data
Check quality of normalized data
Analyze data
Interpret and report findings
Collect mRNA
Reverse transcribe to cDNA
Label sample
+
Microarray
• Summarization
• Identification of differentially expressing genes
• Pattern discovery
• Class prediction

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.