

## 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. Biologists are now confronted with a huge number of genes being newly identified and 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 even 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 its 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. The proper storage, analysis and interpretation of these data have turned out to be a major challenge.

Our focus is on the analysis part. After all, the data alone does not constitute knowledge. It must be first analyzed, 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 has 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 fashion, 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. CDA ensures that chance patterns are not mistaken for real structure. Even at this phase, 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. For this reason EDA techniques have found their way into the world of data mining (Fayyad, Piatetsky-Shapiro, and Smyth, 1996). In data mining, broad-based methods that have the capability 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 be apparent and at challenging assumptions about the data that otherwise may be 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), Tukey (1986), Brillinger, Fernholz, and Morgenthaler (1997), Fernholz, Morgenthaler, and Stahel (2001), and has gained a large following as the most effective way to seek structures in data. Hoaglin, Mosteller, and Tukey (1983) provide an excellent introduction to EDA. Cabrera and McDougall (2002) give a wide range of applications of EDA to real world problems.

That is not to forget CDA. Tukey (1980) argues that exploratory and confirmatory analyses 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, MINITAB, and STATISTICA.

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 few are listed below:

- DNAMR (<http://www.rci.rutgers.edu/~cabrera/DNAMR>), which stands for “DNA Microarray Routines,” is a collection of R and SPLUS programs developed by the authors of this book. 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.
- The Bioconductor project (<http://www.bioconductor.org>), based at the Biostatistics Unit of the Dana Farber Cancer Institute at the Harvard

Medical School and the Harvard School of Public Health, produces open source R software for scientists and statisticians working in bioinformatics, with primary emphasis on inference using DNA microarrays.

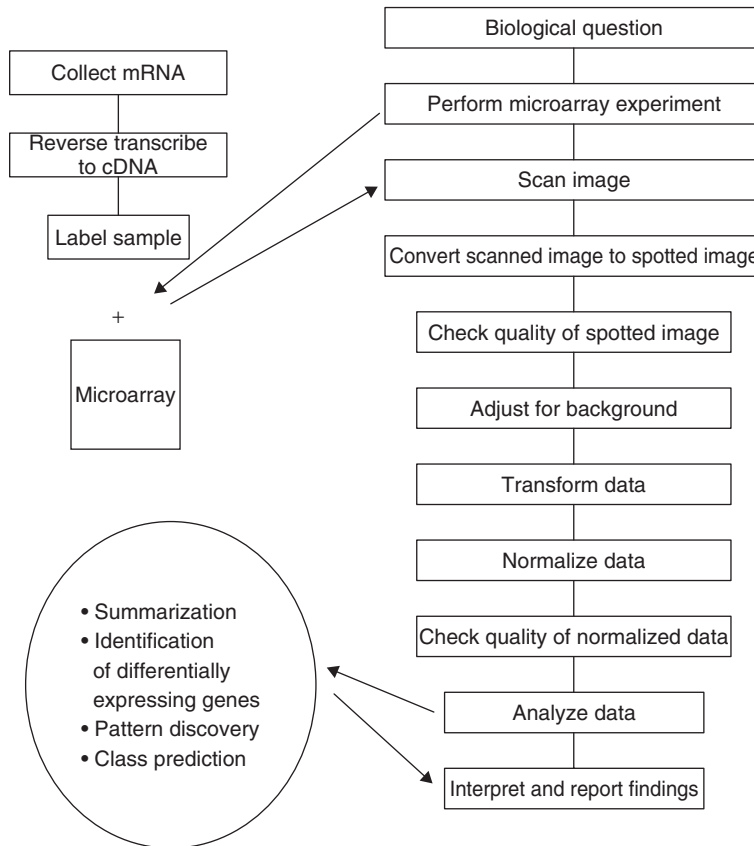
- MA-ANOVA (<http://www.jax.org/research/churchill/software/anova>) is a set of functions written in Matlab by the Statistical Genetics group at the Jackson Laboratory for analysis of variance of microarray data.
- DRAGON (Database Referencing of Array Genes ONLINE) is a series of tools for analyzing and interpreting microarray data that has been developed by a group of researchers at the Johns Hopkins University (<http://pevsnerlab.kennedykrieger.org/index.html>). It has an annotate tool that can be used to add any type of biological information to the lists of genes. The DragonView suite of tools can be used to visualize microarray data relevant to the information derived from the annotate tool. SNOMAD is a collection of R programs for normalizing DNA microarray data.
- The Stanford University Laboratory for the Statistical Analysis of Microarray Data (<http://www-stat.stanford.edu/~tibs/lab>) has software called SAM: Significance Analysis of Microarrays, ScanAnalyze, Cluster, and TreeView.

Although such packages are adequate for routine analyses, for more complex experiments the greater flexibility afforded by software developed in-house may be more desirable. In addition care must be taken that this software is not blindly (mis)used by an individual who does not have enough understanding of the details of the procedures—using the wrong methods to analyze data from an experiment may produce meaningless “findings” or, at the very least, be less than optimal. Unfortunately, few off-the-shelf packages offer a comprehensive data-handling system that integrates all of 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

Exploratory and confirmatory data analysis techniques can be applied to microarray data to:

- Assess the quality of a microarray
- Assess the quality of the individual spots on a microarray
- Determine which genes are differentially expressed
- Classify genes based on how they co-express
- Classify samples based on how genes co-express



**Figure 1.1** Schematic of a typical microarray data analysis.

Following this, the investigator will generally try to:

- Connect differentially expressed genes to sequence databases
- Locate differentially expressed genes on pathway diagrams
- Relate expression levels to other cell-related information
- Determine the roles of genes based on patterns of co-expression.

Often this process will culminate in an insight of interest.

Figure 1.1 shows schematically the path of a typical microarray data analysis. The reader may find it useful to periodically refer to it. In this book we will present a collection of techniques for analyzing microarray data. 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 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. The next two chapters deal with multivariate methods: Chapter 9 discusses unsupervised classification methods and Chapter 10 discusses supervised classification methods. Chapter 11 describes protein arrays. A typical protein array experiment is outlined and methodology for analyzing protein array data is described.

The website <http://www.rci.rutgers.edu/~cabrera/DNAMR> will function as a companion to this book. It contains color versions of the figures, software, updates and any amendments related to the book.