11

Analysis Overview

V. J. Carey and R. Gentleman

Abstract

Chapters in this part of the book address tasks common in the downstream analysis (after preprocessing) of high-dimensional data. The basic assumption is that preprocessing has led to a sample for which it is reasonable to make comparisons between samples or between feature-vectors assembled across samples. Most examples are based on microarray data, but the principles are much broader and apply to many other sources of data. In this overview, the basic concepts and assumptions are briefly sketched.

11.1 Introduction and road map

Chapters in this section address approaches for deriving biological knowledge and formally testing biological hypotheses on the basis of experimental data. We concentrate on DNA microarray data, but other high-throughput technologies such as protein mass spectrometry, array comparative genomic hybridization (aCGH), or chromatin immuno-precipitation (ChIP) are also relevant.

The major focus of this section is on the application of unsupervised and supervised machine learning technologies to the analyses of these large complex data sets. We begin by considering distance measures, as these play an important role in machine learning. We next consider supervised and unsupervised machine learning in some detail and consider multiple testing methodologies and their application to the problems considered in the earlier chapters. The final chapter reviews a Bioconductor approach to browser-based workflow support for downstream analysis.
11.1.1 Distance concepts

It is both common and fruitful to invoke metaphors of spatial organization when discussing high-dimensional data structures arising in various disciplines. Thus while it is sometimes physically sensible to speak of the distance between two genes as a quantity measured in base pairs along a chromosome, it is also sometimes appropriate to speak of the distance between two genes as a quantity measured by the correlation of expression values obtained on a series of samples. The former concept of distance is precise but breaks down for genes present on different chromosomes, whereas the latter concept of distance can be made meaningful in a wide variety of settings. Chapter 12 describes conceptualizations and formalisms of distances for general structures represented in mathematical models of multidimensional spaces. The implications for microarray data analysis are numerous. The definition of a gene cluster in expression space over a series of samples is crucially dependent on selection of a distance definition. Cluster structures and inferences on co-regulation may change when aspects of the underlying distance model are altered. Distances among samples defined in terms of sample phenotype or clinical features are also of interest, but the mathematical construction of a distance function for such features can be complex.

11.1.2 Differential expression

A very common objective of microarray studies is the identification of sets of genes that are consistently expressed at different levels under different conditions. Chapter 14 illustrates this activity with data on leukemia, kidney cancer, and estrogen responsiveness.

11.1.3 Cluster analysis

Identification of shared patterns of expression across samples is basic to exploratory reasoning about co-regulation. Chapter 13 describes new developments in hierarchical clustering based on intensive resampling and evaluation of strength of cluster membership based on the silhouette function. This function, defined formally in Section 13.2.7, measures the relative magnitudes of within- and between-cluster proximities.

11.1.4 Machine learning

The volume of information in high-throughput bioinformatics gives rise to some doubts that traditional approaches to exploratory and confirmatory statistical inference can discover the latent patterns from which new biological understanding can be developed. Machine learning theory and methods