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 attack problems of pattern recognition in voluminous noisy data with minimal human input. Chapter 16 describes basic concepts of computational learning theory and illustrates a number of applications of such tools as neural nets and random forests to microarray data. Chapter 17 specializes the focus to learning procedures based on weighted voting results among ensembles of learners.

11.1.5 Multiple comparisons

Effective use of statistics with data involves recognizing the inherent tradeoff between sensitivity and specificity of inference procedures. In the context of differential expression studies, the sensitivity of a procedure is related to its tendency to identify differential expression when it is actually present. The specificity of a procedure is related to its tendency to refrain from identifying differential expression when it is in fact not present. When large numbers of inferences are attempted, as is common in microarray studies, the calibration of sensitivity and specificity of procedures is complex and requires understanding of statistical dependencies among the test statistics used for inference. Chapter 15 reviews the key concepts and illustrates fundamental tools available in Bioconductor.

11.1.6 Workflow support

Inference with Affymetrix microarray data proceeds from capture of CEL files with minimally processed intensities, specification of sample-level covariates, selection of analysis strategies, calibration of multiple test procedures, and interpretation of resulting gene sets using biological meta-data. The webbioc package (Chapter 18) is a browser-based interface that guides users through these steps.

11.2 Absolute and relative expression measures

One of the main differences between the Affymetrix and cDNA array technologies is that Affymetrix arrays are typically used to measure the *overall* abundance of a probe sequence in a target sample, whereas cDNA arrays typically measure the *relative* abundance of a probe sequence in two target samples. That is, the expression measures for Affymetrix arrays are typically *absolute* (log) intensities, whereas they are (log) *ratios* of intensities for cDNA arrays. In many cases, one of the samples in a cDNA array hybridization is a common reference used across multiple slides and whose sole purpose is to provide a baseline for direct comparison of expression measures between arrays.

For Affymetrix arrays, a direct comparison of expression measures between genes is problematic because the measurement units are not the same across genes. The measured fluorescence intensities are roughly proportional to mRNA abundance, but the proportionality factor is different for each gene. When using short oligonucleotide arrays, it is a function of the probes used and in particular of the frequencies of the different nucleotides in each probe. What we mean specifically is that between-sample, within-gene comparisons are valid and sensible, but within-sample, between-gene comparisons are not easy to make. If for gene X patient i has an estimated expression measure of 100, while for gene Y that same patient has an expression value of 200, these observed data tell us nothing about the real relative abundance of the mRNAs for these two genes. There could, in fact be more copies of the mRNA for gene X. On the other hand, if a second patient, j, say has an expression measure of 200 for gene X, we would conclude that the abundance of mRNA for X in patient j is likely higher than that observed in patient i.

For cDNA arrays, abundance is not measured directly but rather relative to some standard reference. Consider a patient with estimated relative abundance of 1 for gene X and 2 for gene Y. Then, we infer that gene X is expressed at approximately the same level in patient i as in the reference sample, while gene Y has approximately twice the abundance in patient i as in the reference sample. Note that we still do not know if the absolute abundance is the same or not since the we do not know the true abundance of either mRNA in the reference sample.

In some sense, the distinction between the two types of expression measures is artificial, as one could always select a particular Affymetrix array to use as a reference and take ratios of all expression measures to this referent. This will not be quite as successful as it is for cDNA arrays, because with cDNA arrays both the sample of interest and the reference sample are co-hybridized to the same slide. Co-hybridization is a form of blocking, and blocking in experimental design can provide substantial increases in precision.

We would like to stress that whether there is any real difference between the use of absolute and relative measures depends on the distance being considered, as demonstrated below.

A secondary consideration, that will not be explored here, is that differences in the probes can affect the variability of the expression measures. For short oligonucleotide arrays, there is some evidence that the variability of the estimated expression levels can be quite different across probes for the same gene. The variability may be a function of the mean level of expression (Rocke and Durbin, 2001), but there can be other substantial sources of variation as well. It is unlikely that this phenomenon is restricted to short oligonucleotide data. In particular, one would expect cDNA array data to exhibit similar behavior. Taking ratios with respect to a common reference (that was subjected to the same hybridization and scanning conditions) may provide some relief.