Abstract
In many gene expression studies, the goals include discovery of novel biological classes and identification of genes whose expression can reliably be associated with these classes. Here we present a statistical analysis approach to facilitate both of these goals. The key idea is to model gene expression using latent categories that can be interpreted as a gene being turned “on” or “off” compared to a baseline level of expression. This three-way categorization is used for defining a reference in the unsupervised setting, for removing noise prior to clustering, for defining molecular subclasses in a way that is portable across platforms, and for defining easily interpretable probability-based distance measures for visualization, mining, and clustering.

16.1 Introduction
Molecular class discovery can be defined as investigating a set of previously unclassified subjects from a population to find subgroups that share similar molecular profiles (Duggan et al., 1999). The most common application to date is to data consisting of sets of tumors that are not distinguishable morphologically (Golub et al., 1999; Alizadeh et al., 2000; Bittner et al., 2000; Bhattacharjee et al., 2001; Garber et al., 2001). Subclasses could be defined using genome-wide profiles, but it is often more practical to identify a set of genes that show differential expression across classes. This allows for biological characterization of the classes, for out-of-sample prediction across platforms, and for confirmatory work using alternative gene expression measurement techniques such as RT-PCR. In describing genomic variability at the population level, it can also be important to de-
fine subclasses in terms of differential expression of these genes compared to the typical level in the population.

This chapter describes a statistical approach and software tools to support this type of unsupervised molecular classification. The tools were developed with cancer-related applications in mind but are more generally applicable. Our starting point is a probabilistic definition of differential expression for the unsupervised setting. This is used to define molecular profiles and assess quantities of potential use in classification, such as the probability that an individual belongs to a given profile and the probability that a group of individuals has the same given profile. The detailed context, motivation, and application to the classification of ductal breast cancers using this approach are given by Parmigiani et al. (2002). Successful molecular classification so far has been characterized by a combination of visualization, formal quantitative analysis, and informal a priori information on gene function. Our analysis tools are not intended to replace, but to facilitate, the extensive manual work required to develop novel subclasses.

Exploration of molecular profiles using the tools described in this chapter proceeds along the following lines.

- Identification, for each gene, of a baseline or modal class, using statistical modeling.
- Estimation, for each gene-subject combination, of whether gene expression is at baseline, overexpressed, or underexpressed and derivation of probabilities of these differential expressions. These probabilities can offer an effective way to stabilize the measurements by eliminating a large portion of the noise and the hard-to-cluster variation associated with extreme gene expression values. At the same time, they provide an interpretable scale for classification of tumors to patterns.
- Mining for candidate genes to predict classes using the probability that they follow a pattern of specified characteristics. Mining produces potentially overlapping groups of genes with similar patterns.
- Visualization of probabilities that subjects belong to a specified profile, defined in terms of a small set of interpretable genes. Not all samples need to be assigned to a class.

Once a small set of $s$ genes is identified, a molecular profile can be defined as an $s$-dimensional vector, with each coordinate taking one of three values (overexpressed, underexpressed, or expressed as baseline). This allows for a crisp and portable definition of a molecular profile. Also, in some datasets, the definition of a baseline or modal class allows us to model, and potentially remove, a noise component.

Model estimation is implemented using a Bayesian approach with Markov chain Monte Carlo (MCMC) evaluation of posterior quantities. We describe R software to carry out the estimation. The functions for performing the analysis require only a data matrix of expression values