'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns

ABSTRACT

The gene shaving method is a potentially useful tool for exploration of gene expression data and identification of interesting clusters of genes worth further investigation.

INTRODUCTION

Background Through the use of recently developed DNA arrays, it is now possible to obtain accurate, quantitative (relative) measurements of a large proportion of the mRNA species present in a biological sample. DNA arrays have been used to monitor changes in gene expression during important biological processes (for example, cellular replication and the response to changes in the environment), and to study variation in gene expression across collections of related samples (such as tumor samples from patients with cancer). A major challenge in interpreting these results is to understand the structure of the data produced by such studies, which often consist of millions of measurements. A variety of clustering techniques have been applied to such data, and have proved useful for identifying biologically relevant groupings of genes and samples. Although the underlying principles and computational details of these methods differ, they share the goal of organizing the elements under consideration (such as genes) into groups (clusters) with coherent behavior across relevant measurements (such as samples). Generally absent is any consideration of the nature of the coherent variation. For example, one might want to identify groups of genes that have coherent patterns of expression with large variance across samples, or groups of genes that optimally separate samples into predefined classes (such as different clinical response groups in tumor samples). Here, we introduce a new statistical method, which we call gene shaving, that attempts to identify groups of elements (genes) that have coherent expression and are optimal for various properties of the variation in their expression. Figure 1 shows the dataset used in our study, which consisted of 4673 gene expression measurements on 48 patients with diffuse large B-cell lymphoma (DLCL). These data have been described in detail previously. The column labels refer to different patients, and the rows correspond to genes. The order of rows and columns is arbitrary. Some authors have recently explored the use of clustering methods to arrange the genes in some systematic way, with similar genes placed close together (see for developments and for an overview). In Figure 2, we have applied hierarchical clustering to the genes and samples separately. Each clustering produces a (non-unique) ordering, one that ensures that the branches of the corresponding dendrogram do not cross. Figure 2 displays the original data, with rows and columns ordered accordingly. Some structure is evident in Figure 2, and this method can be used to recognize relationships among the genes and samples.With any method that reduces the dimension of the data, however, finer structure can be lost. For example, suppose the expression of some subset of genes divides the samples in an informative way, correlating with the rate of proliferation of tumor cells, for example, whereas another subset of genes divides the samples a different way, representing the immune response, for example. Then methods such as two-way hierarchical clustering, which seek a single reordering of the samples for all genes, cannot find such structure. The method of gene shaving we describe here is designed to extract coherent and typically small clusters of genes that vary as much as possible across the samples. Figure 3 shows three gene clusters for the

DLCL data, found using shaving. Some of the genes within each cluster lie close to each other in the hierarchical clustering of Figure 2, but others, and the clusters themselves, are quite far apart. In Figure 3 the samples have been ordered by values of the average gene expression. This average gene is a good representative of the cluster, as all the members are so similar. The variance measures at the top of each cluster are discussed in more detail later. The clusters are all of different sizes. We use an automatic method for determining the size of the clusters, based on a randomization procedure that protects us from looking too hard in the large sea of genes and finding spurious structure. The three cluster-average genes, one from each cluster, are reasonably uncorrelated (see below and Figure 6). This is another aspect of the shaving process - it seeks different clusters, where difference is measured by correlation of the cluster mean. Figure 4 shows the results of a hierarchical clustering applied to the three column-average genes. Whereas hierarchical clustering suggests two main gene groupings, the shaving process may suggest more useful groupings. This article is organized as follows. In the section 'Gene shaving' we describe the method itself. The section entitled 'The gap estimate of cluster size' outlines the gap test for choosing the cluster size. In the section 'Predicting patient survival' we try to predict patient survival from gene cluster averages. 'Supervised shaving' is discussed in the following section. Finally, in the 'Conclusions' we propose some further generalizations. A more statistical treatment of gene shaving is given in.

CONCLUSION

Conclusions We have proposed a set of 'shaving' methods for isolating interesting clusters of genes from a set of DNA microarray experiments. The methods may be unsupervised, or may be supervised - that is, use information available about the samples such as a class label or survival time. The proposed shaving methods search for clusters of genes showing both high variation across the samples, and coherence (correlation) across the genes. Both of these aspects are important and cannot be captured by simple clustering of the genes, or thresholding of individual genes based on the variation over samples. With our model-based approach for supervised shaving, one can incorporate other prognostic factors in the search for interesting gene clusters. If an outcome such as survival time is available for each sample, the method searches for a gene cluster whose column average gene has a significant effect, possibly the presence of other prognostic factors, for predicting the outcome. The microarray data xij we have considered are real-valued expression levels. However, other kinds of arrays produce different kinds of data. In particular, some arrays detect the presence or absence of single-nucleotide polymorphisms (SNPs), so that the xij values take on one of $k \geq 2$ unordered values. The shaving methods described can be easily modified to handle this kind of data. In detail, we construct k data matrices X1, X2 ... Xk, each of size $n \times m$. The ijth element of Xj is 1 if xij falls in class j, and zero otherwise. Letting $\sum$jj = 1, 2, ... k be the $n \times n$ covariance matrices of the genes in each Xj, we simply apply principal component shaving, using $\sum$j, as the variance matrix for the penalty. This can be done unsupervised, or a supervision term can also be added.