Supervised harvesting of expression trees

ABSTRACT

The use of tree harvesting can enable the analysis of gene expression data and the discovery of intriguing clusters of genes that require further investigation.

INTRODUCTION

The use of DNA arrays has enabled the quantitative quantification of a significant portion of biological mRNA species, providing an overview of the current situation. The use of DNA arrays has enabled the observation of gene expression changes during crucial biological processes, such as cellular replication and environmental changes, as well as the examination of differences in gene Expression among related samples, like cancer patients' tumor samples. Understanding the structure of the data generated by such studies, which frequently contains millions of measurements, can be a difficult task when dealing with this type of analysis. A variety of clustering techniques have been used to analyze these data, and they have also been useful in identifying biologically relevant groupings of genes and samples. The methods' underlying principles and computational details vary but they both aim to organize the elements under consideration (such as genes) into groups that exhibit a coherent behavior across relevant measurements (like samples), without taking into account the nature of the coherent variation. Identifying genes in groups with consistent expression patterns across samples and a clear separation of samples into predetermined classes (e.g. different clinical response groups in tumor samples) could be useful, for instance. A new statistical technique called gene shaving is used here 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 we used, which included 4673 gene expression measurements on 48 patients with diffuse large B-cell lymphoma (DLCL). Detailed descriptions of the data have been provided previously. The column labels indicate different patients, while the rows and columns indicate genes. There is no specific order of these data, and some authors have explored clustering to arrange the genes in a systematic manner, with similar genes placed close together (see developments for more information). We have used hierarchical clustering to classify the genes and samples separately, resulting in a non-unique ordering that prevents branch crossing. Figure 2 illustrates the original data with rows and columns. Figure 2 displays a certain level of structure, which can be utilized to identify connections between the genes and samples. Any approach that decreases the data size can result in the loss of more detailed structure. Let's say that the division of samples is caused by the expression of one subset of genes, which corresponds to the proliferation rate of tumor cells, while another subgroup, representing the immune response, does not. The use of two-way hierarchical clustering and shaving (gene shaving) is not an option because it does not find a consistent pattern for all genes. Instead, the technique involves extracting small, coherent clusters of genes that vary as much across the samples. The hierarchical clustering in Figure 2 shows some genes located close to each other, while others are quite far apart. In Figure 3, the samples are arranged according to the average gene expression values, which is a good indicator of the cluster's similarity. The variance measures at the top of each cluster are outlined later. They are all different sizes. We employ an automatic method to determine the size of the clusters, which prevents us from perceiving the genome as spurious structure by combing through

thousands of genes. Figure 6 illustrates that the three cluster-average genes, with one from each cluster, are not significantly correlated. As a result, they participate in the shaving process to identify distinct clusters, measuring the difference by using the correlation of the cluster mean. Figure 4 displays the results of hierarchical clustering based on the expression of these genes. The shaving process can provide alternative useful gene groupings, unlike hierarchical clustering. Our section on 'Predicting patient survival' endeavors to use gene cluster averages as a reference point for forecasting the patient's survival. We discuss subsequently supervised shaving in the following section and suggest some additional generalizations in our section called, in particular, "Conclusions." A more detailed analysis of gene shaving is available in later sections with different appendages.

CONCLUSION

This exploratory study demonstrates "...any association between the polymorphisms in codon 27 of ADRB2 and in [ADRB3] genes that may be associated with increased risk of breast cancer"; however, additional studies across larger samples and/or across different ethnicities are needed to further explore this effect.