Applying Classification Separability Analysis to Microarray Data

Zhen Zhang*, Grier Page*, and Hong Zhang
* Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC
Informatics, Charleston, SC
Barnhill Genomics, Savannah, GA

ABSTRACT

The majority of published papers with microarray based gene expression data analyses use some variations of clustering algorithms to organize the expression data so that genes sharing similar expression profiles across multiple experiments are arranged together for easy identification of patterns. In a similar fashion, individual experiments with similar expression profiles over the entire set of genes may also be clustered together. The main advantage of this approach is that it provides a holistic view of expression patterns across the entire set of observations. A noticeable drawback of this approach is, however, that the majority of genes in the dataset, some of them even with strong expression variation patterns, might not be associated at all with a particular end point of interest (e.g., different phenotypes or experiment conditions). As a consequence, those expression patterns that are truly associated with the end point would have to have a strong presence in terms of number of genes of similar profiles so that they could be detected among the large number of non-contributing expression patterns. It has recently be suggested under the name of singular value decomposition that expression data be projected to the so-called eigengene x eigenarray space with a much small dimensionality for better data interpretation. The approach is similar to principal component analysis in which the major eigengenes correspond to the direction represented by genes with the largest variance in their expression levels. Again, a drawback of this method is that those genes with a large variance in expression levels across different experiment arrays might not necessarily be associated with the end point of interest in the analysis.

In this paper, we present a new approach of iteratively sorting genes according to their collective contribution to the separation of different values in an end point of interest (e.g., specimens from breast tumor tissues or normal breast tissues). After sorting, those genes with a significant collective contributions are rearranged using a clustering algorithm for visualization and pattern detection. As an example, we applied this new approach to the analysis of the budding yeast saccharomyces cerevisiae data with the purpose of identifying genes that response to heat shock. The end point here therefore was chosen to be the separation of experiments of yeast through normal division cycles and those under heat shock at different time points. We were able to show that the first 500 sorted genes, after clustering, demonstrated expression profile patterns over a large number of genes that are strongly associated with the heat shock stress. Such patterns decrease progressively towards the end of the sorted genes, with the last 500 genes showing almost completely random patterns after clustering (figure 1). Our method yielded clusters of genes whose response to heat shock seem to be biologically plausible (figures 2 and 3). One down regulated cluster is composed primarily of ribosomal proteins. This is expected since protein synthesis is strongly down regulated in heat shock conditions. The cluster of strongly up regulated genes appears to have two classes of genes. One is the class of heat shock proteins, chaperonins, and other genes involved in cellular and protein stability. The other class of genes is those involved in metabolism and the uptake of metabolites. Thus in response to heat shock the cells are not making many new proteins and at the same time are trying to protect the existing proteins and to expand the list of possible energy sources.

Figure 1, Clustering results of sorted genes, in groups of 500 each, according to their collective contribution to the separation of experiments of yeast under normal division cycles and those under heat shock stress at different time points. In this figure, the vertical lines corresponds to the genes and the horizontal lines are the experiments (arrays).

Figure 2, A group of up-regulated genes associated with yeast under heat shock at different time points.

Figure 3, A group of down-regulated genes associated with yeast under heat shock at different time points.