Analysis of Gene Expression Data with GeneSpring and MetaMine.

Andrew Conway

The expression data collected by Golub and coworkers (Golub et al., 1999, Science v286, p531-37) and of Spellman and coworkers (Spellman et al., 1998, Mol. Biol. Cell. v9, p3273-97) were analyzed using the GeneSpring™ software package. This software is an integrated workbench that facilitates the rapid discovery of biological expression patterns. In these studies, we demonstrate that the tool is particularly useful for both class discovery and class prediction. The two data sets of were imported, normalized and analyzed using GeneSpring™.

Perhaps the biggest benefit of the analysis techniques we will show is the convenience, speed, ability to cross reference, and lack of computer or statistical expertise needed by the user. Both data sets can be easily analyzed using mature commercially available and supported software, finding similar results to both Golub and Spellman in minutes, as well as some other results. Analyses of one form can then be simply cross referenced to other types of analyses. Some of our results were even obtained without human direction by our autonomous data mining tool MetaMine™, and the full and detailed reports generated will be shown. We can demonstrate these analyses and cross references in real time.

Using the independent set of 38 leukemia patients studied by Golub et al., we were able to successfully construct a prediction method for 34 of 38 patients, which discriminates perfectly between the two leukemia classes (myeloid vs. lymphoblastic). Genes were evaluated as potential class predictors based on their individual ability to distinguish between the classes in the training set. For each gene, all possible threshold expression values were considered, and the ability of that gene to discriminate between the classes was determined. The genes were then ranked according to their best discrimination score. The top ranking genes were then used to classify samples in the independent data set by considering the known classifications of their k nearest neighbors in the training set.

The predictor set of 100 human oncogenes, which was highly successful in discriminating between leukemia classes, was then divided into two subgroups: one group whose expression was elevated in ALL patients, and a second group whose expression was attenuated in ALL patients. Like Golub and coworkers, we found that both of our predictor lists were rich in genes related to the cell cycle, to protein synthesis, and to protein trafficking. To model the manner in which these genes might behave in the leukocyte cell cycle, we chose to examine the expression patterns of homologous genes in Saccharomyces cerevisiae. For this purpose, we found the expression data of Spellman and coworkers to be very useful, as was the use of the clustering and visualization tools available in GeneSpring™. A list of yeast genes was constructed whose members were either homologous to, or were functional orthologs of members in our predictor set. We show that discrete groups of these genes are differentially co-expressed in the yeast cell cycle and appear to have common regulatory elements.