A Ranking Method to Improve Detection of Disease Using Selectively Expressed Genes in Microarray Data.

Virginie Aris and Michael Recce,
Center for Applied Genomics, MSB 901-A, 185 South Orange Ave, Newark, NJ 07103, and Center for Computational Biology and Bioengineering, NJIT, University Heights, Newark, NJ 07102.

Cancer classification can be successfully implemented through expression profiling with DNA microarrays. The aim of the Golub et al. study (1999) was to classify and predict classes of Leukemia by determining which genes have expression levels that are most correlated with different disease states. Their analysis method was selective for genes that have above background expression in all of the subjects. While this is true for many genes, others are selectively expressed in one of the disease states, and are not distinguishable from the background in the other states. We describe a method to improve classification of the data in the Golub et. al. through analysis of these selectively expressed genes. Further this method reduces reliance on scaling or normalization factors in the comparison of data across subjects.

Several genes were in the Golub et. al. data set are selectively expressed between acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). We show that the presence or absence of expression in the ten most selective genes, represented as a binary number (expressed =1, not expressed =0) is sufficient to correctly classify and diagnose the disease state of the subjects in the data set. In this initial analysis the level of gene expression is not used. The exemplar, or cluster center, for each of the two diseases is computed as the real-valued average of the (expressed/not expressed) binary values for each of the most selective genes of the subjects with each disease (27 ALL, 11 AML). Then the Euclidian distance is computed for each subject to each of the two exemplars. Members of a cluster are closer (smaller distance) to the exemplar of that cluster than to the exemplar of the other cluster. For example the range of distances from the AML subjects to the AML exemplar is 0.05 to 0.28, and the range of distances of this group to the ALL exemplar is 0.62 to 0.87. These data, along with the distances of the ALL subjects from the two exemplars show that the two clusters are well separated, with no overlap (false positives or false negatives).

Certainly additional information is present in the level of gene expression, but normalization or scaling errors across subjects or slides make it difficult to determine the precision of these numbers. The number of genes that have above background expression levels vary across subjects, and there is a high (0.47) linear correlation between this number and the inverse of the scaling factor that was used to adjust the expression level. In other words, slides that have fewer genes with above background expression levels require larger scaling. There is less uncertainty in the comparison of expression levels on an individual slide. The relative expression levels on a slide can be used to rank the genes, and this ranking can be used to reduce the variation in the number of active genes. We show that there is an optimum level of application of this ranking method. Further it significantly improves the separation of the two clusters, as measured by the changes in the distribution of distances between the cluster members and the cluster exemplars.

We show that this method can also be used to distinguish between the two sub-types of ALL described by the Golub et al. study, and can be used to diagnose the disease present in the second data set. This method is complimentary to the one used in the prior study, since it focuses on the genes that are selectively expressed, rather than on differential expression levels. A synthesis of the two methods should improve the classification performance, since these two sets of genes are largely independent.