Discovery: A Tool for Class Prediction Using Gene Expression Data

Louxin Zhang, Zhuo Zhang, and Song Zhu

Kent Ridge Digital Labs & BioInformatics Centre, Singapore 119613
Email: zzhang, lxzhang, zhusong@krdl.org.sg.

As a part of our package for the analysis of gene expression data, we implement a tool for class prediction in CGI, Java, Perl and C-languages. The problem of class prediction is roughly like diagnosis: given a set of known classes, determine the correct class for a new sample. This tool for class prediction is based on an elegant method proposed in a recent work (Slonim *et al.*, 2000). Because of its simplicity, the method has great potential in improving cancer classification and diagnosis using gene expression data. In fact, Golub *et al.* (1999) applies the method to human acute leukemia as a test case. An automatically derived predictor is able to determine whether a new leukemia case is acute myeloid leukemia (AML) or acute lymphoblastic leukemia (ALL) with high accuracy. Although the method can possibly be extended to construct multi-class predicator, we just implement it for predicting membership in one of just two classes.

If a class distinction is highly predictable (which is tested as the first step), the program selects a certain number(20, 25 or 30) of genes that are mostly correlated with each class and then determine the correct class for a new sample by the sum of the weighted votes cast by all selected genes (Slonim *et al.*, 2000).

To make the tool easy for use, we use the web interface like other bioinformatics tools. The web brower window is partitioned into four frames as illustrated in Figure 1. The whole process runs in three steps as listed in *left-top frame*. Input fields and parameters appear in the *left-bottom frame*; parameters and concepts are explained in the *right-top frame* so that one can easily check for their definitions if necessary; and outputs appear in the *right-bottom frame*. Since parameters and outputs appear in different frames, the user can view the results while keeping the parameter settings unchanged. This allows one to modify parameter values until one is satisfied with the analysis.

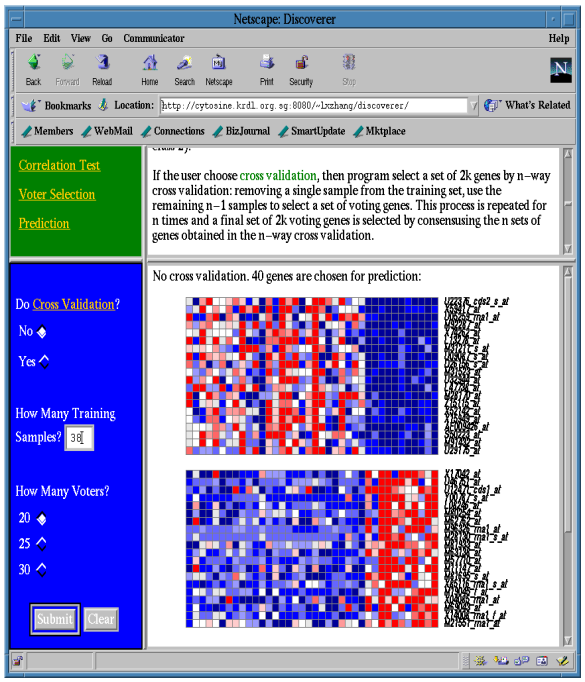

Figure 1: *Configuration of Discovery. Here the voting genes selected in the second step are listed in the right-bottom frame.*

References

D. K. Slonim, et al. Class Prediction and Discovery Using Gene Expression Data, *Recomb'2000*, 262-271, Tokyo, 2000.

T. R. Golub, et al. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring, *Science*(286), 531-7, 1999.