SUPPLEMENTAL MATERIAL FOR:

PITFALLS IN THE USE OF DNA MICROARRAY DATA FOR DIAGNOSTIC AND PROGNOSTIC CLASSIFICATION

Richard Simon[1], Michael D. Radmacher[2], Kevin Dobbin[1] and Lisa M. McShane[1]

[1] Biometric Research Branch, National Cancer Institute, NIH, Bethesda, MD

[2] Departments of Biology and Mathematics, Kenyon College, Gambier OH

GENERATION OF SIMULATED DATA SETS

Each simulated data set is composed of 20 gene expression profiles, each profile consisting of 6000 gene expression measurements. Measurements are in the form of log-expression ratios; the log-ratios of the 6000 genes that compose a single gene expression profile are independent and identically distributed from the standard normal distribution (i.e., with a mean of zero and variance of one). Data sets were generated using Compaq Visual Fortran (© 2000, Compaq Computer Corporation) and the IMSL routine RNNOA for generating standard normal random variables (Visual Numerics, Inc., 1997). Since the gene expression profiles of all specimens in each data set are generated from the same distribution, any distinction found between two subsets of the specimens is spurious. Nonetheless, we performed class prediction on each simulated data set to see if we could accurately predict class membership based on an arbitrarily chosen separation of the data set into two classes of 10 specimens each.

CLASS PREDICTION IN THE SIMULATION

The method of prediction we used was compound covariate prediction (Radmacher *et al.*, in press). First, we performed a univariate, two-sample t-test on each gene to establish the genes that best discriminated between the two classes of specimens. The 10 genes with most significant t-statistics were included in the predictor (these genes are referred to as the differentially expressed genes). The predictor was a compound covariate: a linear combination of the log-expression ratios of the differentially expressed genes. The two-sample t-statistic of each differentially expressed gene served as its weight in the compound covariate. The value of the compound covariate for specimen *i* was:

$$c_i = \sum_j t_j x_{ij},$$

where t_j is the t-statistic for the two group comparison of classes with respect to gene j, x_{ij} is the log-ratio measured in specimen i for gene j and the sum is over all differentially expressed genes.

The last step in prediction was the definition of a classification rule; a simple threshold rule was used. We computed $C_t = (C_1 + C_2)/2$, where C_1 and C_2 are the mean values of the compound covariate for specimens in the training set with class label 1 and class label 2, respectively. In other words, the threshold, C_t, is the midpoint of the means of the two classes. With this classification threshold, a specimen was predicted to be of class 1 if its compound covariate was closer to C_1 and to be of class 2 if its value was closer to C_2. Compound covariate prediction is implemented in BRB ArrayTools, a visualization and statistical analysis software package for microarray gene expression data available for download at http://linus.nci.nih.gov/BRB-ArrayTools.html.

CROSS-VALIDATION IN THE SIMULATION

The steps in prediction were:

1. Gene selection (10 most differentially expressed genes based on two-sample t-test).
2. Computation of a weight for each gene (univariate two-sample t-statistic).
3. Computation of prediction rule (classification threshold, C_t, as defined above).

Different levels of leave-one-out cross-validation were implemented in the class prediction. They were:

1. Resubstitution (no cross-validation).
2. Cross-validation after gene selection (removal of left out specimen between steps 1 and 2 above).
3. Cross-validation prior to gene selection (removal of left out specimen before step 1).

REFERENCES

Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, *et al.* Molecular classification of cancer: class discovery and class prediction by gene expression profiling. Science 1999; 286:531-537.

Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002; 9:505-512.