Technical Report 001
July 1, 2001

A Paradigm for Class Prediction Using Gene Expression Profiles

Michael D. Radmacher, Lisa M. McShane and Richard Simon

Affiliation of authors: Biometric Research Branch, National Cancer Institute, 6130 Executive Boulevard, Bethesda, MD 20892-7434.

Correspondence to: Michael D. Radmacher, Ph.D., National Cancer Institute, 6130 Executive Blvd., MSC 7434, Bethesda, MD 20892-7434. *Phone*: 301-496-3307. *Fax*: 301-402-0560. *E-mail*: mdradmac@helix.nih.gov

Abstract

We propose a general framework for prediction of pre-defined tumor classes using gene expression profiles from microarray experiments. The framework consists of 1) evaluating the appropriateness of class prediction for the given data set, 2) selecting the prediction method, 3) performing cross-validated class prediction and 4) assessing the significance of prediction results by permutation testing. We describe an application of the prediction paradigm to gene expression profiles from human breast cancers, with specimens classified as positive or negative for BRCA1 mutations and also for BRCA2 mutations. In both cases, the accuracy of class prediction was statistically significant when compared to the accuracy of prediction expected by chance. The framework proposed here for the application of class prediction is designed to reduce the occurrence of spurious findings, a legitimate concern for high-dimensional microarray data. The prediction paradigm will serve as a good framework for comparing different prediction methods and may accelerate the development of molecular classifiers that are clinically useful.

Microarray technology has made possible the detailed molecular characterization of tumor specimens through the simultaneous measurement of mRNA expression levels for thousands of genes. These gene expression "profiles" (as they are frequently called) can be used to discover biologically similar groups of tumors not previously recognized, to identify genes differentially expressed between two or more predefined groups, or to construct multivariate predictors of group membership. This paper addresses statistical considerations for the last of these goals, tumor class prediction. Examples include predicting whether or not a tumor will respond to a particular chemotherapy, predicting if a patient will be a long-term or short-term survivor, and predicting which tumors possess a particular gene mutation.

The Class Prediction Paradigm

We propose a general framework for prediction based on gene expression profiles. The framework consists of 1) evaluating the appropriateness of class prediction for the given data set, 2) selecting the prediction method, 3) performing cross-validated class prediction and 4) assessing the significance of the prediction results by permutation testing. This model is designed to address the complexities of analyzing high dimensional data and to reduce the detection of differences between classes that are spurious.

Appropriate setting for class prediction. "Supervised" analyses are useful when, in addition to gene expression profiles, we have supplemental information about each specimen and want to study relationships between this information and expression profiles. Specifically, with class prediction the goal is to predict the supplemental information of a specimen from its expression profile. The information to be predicted may be the patient's response to therapy, survival outcome or mutation status, and we refer to this information as a "class label". Class labels may also be thought of as phenotypes. In this paper we consider the situation where class labels are dichotomous (e.g., poor and good responders), but the method is generalizable to more than two classes.

For problems where the class labels are specified in advanced, supervised methods may be more powerful than unsupervised methods (e.g., cluster analysis). In clustering specimens, global similarity metrics are often used which represent similarity among specimens with regard to all of the genes represented on the microarray. Differences between classes that manifest themselves with regard to a small number of genes may not have much influence on global similarity metrics and therefore result in clusters that have little association with class labels. With supervised methods, genes that distinguish the classes are specifically identified. Such methods are more appropriate for discerning differences between classes but carry the risk of identification of spurious differences between the classes. Hence, the proper analysis of such methods is more complex.

The proper setting for use of class prediction methods is one in which class labels are known a priori: the labeling of specimens should not be based on their expression profiles. For example, the class labels used in prediction should not be derived by performing a cluster analysis on the same specimens. Classification performed in such a way will certainly be biased in favor of good prediction for the data set being studied because the class labels were defined by their ability to, in some sense, optimally separate the specimens of this particular set into two groups.

Selection of a prediction method. A myriad of prediction methods exist (e.g., classification and regression trees, discriminant analysis, stepwise logistic regression). However, prediction from gene expression data presents its own unique challenges. Not the least of these is that the number of covariates (i.e., genes) on which measurements are made vastly exceeds the number of specimens to classify. Thus, the risk of over-fitting the data is great in microarray studies: a complex multivariate analysis may very well result in a perfect discriminator for the data set at hand, but perform poorly on independent data sets. Prediction methods intended for situations in which there is a small number of covariates and a large number of observations may thus be poor choices for molecular classification of specimens. Predictors built using expression data must be restricted in complexity, with the competing demands of simplicity and accuracy properly balanced (1).

Several prediction methods have been utilized in the analysis of microarray data. Golub et al. constructed a "weighted voting method" and used it to distinguish between two types of human acute leukemias (2). Support vector machines were used both for the categorization of genes into functional classes and the prediction of ovarian tissue specimens as normal or cancerous (3,4). Ben-Dor et al. built a clustering-based classification technique and applied it to various data sets, including those analyzed with the above two methods (5). We developed a method, compound covariate prediction, that was used to predict the BRCA1 and BRCA2 mutation status of breast cancer specimens

(6). This method is described in detail below and is used to demonstrate the proposed paradigm for class prediction.

Regardless of the method chosen for prediction, a means of assessing the result is necessary. Good prediction accuracy is an obvious criterion but, as noted above, must not represent an artifactual over-fitting of the data. Ultimately, we desire a method that will accurately predict class labels for new specimens that were not involved in the creation of the prediction rule. Two steps can be taken, even in the absence of an independent validation set, to assess the quality of a prediction rule: cross-validation and permutation testing. These steps are discussed below.

Leave-one-out cross-validated class prediction. An important component of the class prediction framework is cross-validation. Cross-validation of prediction results is accomplished by leaving out a portion of the data, building the prediction rule on the remaining data (referred to as the training set) and predicting the labels of the left out data. In doing so, it is important to treat selection of genes for inclusion in the predictor as a step in the building of the prediction rule and, hence, subject to cross-validation. Leave-one-out cross-validation is a logical choice for relatively small sample sizes, with each specimen left out one at a time and the label of the left out specimen predicted from the remaining data (7,8), and is the method we use for microarray analysis.

The importance of cross-validation is demonstrated by the following simulation example. Suppose we have data for 14 microarray experiments using a 6000-gene cDNA microarray (each experiment involving a unique specimen and a common reference sample). Further assume that for each experiment, log-expression ratios for each gene are independent and identically distributed from the standard normal distribution. Two thousand such 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 a 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 discriminate between the first 7 specimens and the last 7 specimens (labeled 'Class 1' and 'Class 2', respectively) in each data set. The method of prediction we used was compound covariate prediction based on the 10 most differentially expressed genes (this method is described in detail below), and we implemented both a non-cross-validated and leave-one-out cross-validated scheme.

Non-cross-validated prediction of class labels was perfect in nearly all of the simulated data sets; the maximum number of misclassified specimens in a data set was 1 (Table 1). Hence, without cross-validation it was always possible to build an excellent class predictor for the data, even though the gene expression profiles for the two classes were generated from the exact same distribution. The reason for this is that in a non-cross-validated setting, the gene expression profiles of the specimens whose labels are to be predicted are included in the construction of the prediction rule. This leads to over-fitting of the data and an optimistically biased estimate of the error (7-9). Leave-one-out cross-validated prediction corrects for much of this bias: the median number of cross-validated misclassifications for simulated data sets was 8 and the maximum number was 14 (Table 1). Thus, cross-validation provides a more realistic estimate of the misclassification error rate one can expect when applying the predictor to independently collected data.

Permutation testing. Cross-validation is an important aspect of the class prediction process but is not sufficient for assessing the significance of a classification result: a small cross-validated error rate does not guarantee that the subclassification of specimens is authentic. Because of the high dimensionality of gene expression data, it may be possible to achieve relatively small error rates even for random data. For example, in the simulation described above more than 6 % of the data sets resulted in 2 or fewer cross-validated misclassifications (Table 1). We use a permutation test to assess the significance of our cross-validated error rate. The idea underlying the permutation test is this: under the null hypothesis that no systematic difference in gene expression profiles exists between the two classes, it can be assumed that assignment of gene expression profiles to class labels is purely coincidental (10). This situation can be mimicked by randomly permuting labels among the gene expression profiles. Considering every possible permutation of the class labels among specimens, the proportion of the permuted data sets that have a misclassification error rate less than or equal to the observed error rate serves as the achieved significance level in a test against the null hypothesis.

Ideally we would perform an exact permutation test, examining every possible permutation of the class labels. However, in practice this is burdensome due to the large number of permutations for even modest sample sizes. We instead implement a Monte Carlo method, examining *k* randomly chosen permutations and using the proportion of these *k* permuted data sets that result in the same or a smaller number of cross-validated misclassifications as an estimate of the achieved significance level. *k* can be chosen such that the variability in the estimate of the achieved significance level is less than some acceptable amount (11); we set *k* equal to 2000 to ensure that our estimate varies by less than 10 % from a true achieved significance level of 0.05. We reject the null hypothesis if the estimate is less than 0.05.

Application of the Prediction Paradigm using the Compound Covariate Predictor

Here we give an example of class prediction in the setting of paradigm outlined above. First, the motivation for and details of the prediction method we use (i.e., compound covariate prediction) are discussed.

Motivation. Over-fitting of data is a legitimate concern when analyzing microarray experiments but is not unique to the analysis of gene expression data. Tukey proposed the use of compound covariates in a clinical trials setting in which a researcher wishes to relate several dozen covariates to survival in a few hundred patients (12). A compound covariate is a linear combination of the basic covariates being studied, with each covariate having its own coefficient or weight in the linear combination. The microarray setting is far more extreme than what Tukey envisioned—typically thousands of covariates (genes) and a few to several dozen patients. Nonetheless, we have found the analysis of compound covariates to be extremely useful for class prediction using microarray data.

Methodology. Applying the compound covariate method to microarray data, we first reduce the dimensionality of the analysis by selecting a subset of genes for further study (subset selection of genes is a common property of the classification methods that have been applied to microarray data). In the context of class prediction involving two classes, we perform a two-sample t-test and select genes with log-expression ratios that best discriminate between the two groups in the training set. As a selection criterion, we could either predetermine the number of genes, *n*, to include in the subset and choose the *n* genes with largest t-statistics (in absolute value) or predetermine a significance level, α, that genes must meet for inclusion in the subset. The genes included in the predictor are referred to as differentially expressed genes.

With the differentially expressed genes determined, we turn to construction of a single compound covariate that will be used for class prediction. The two-sample t statistic of each differentially expressed gene serves as its weight in the compound covariate. Thus, the value of the compound covariate for specimen *i* is

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

where t_j is the *t*-statistic for the two group comparison of labels 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. This predictor is similar to the weighted voting method of Golub et al. (2) since each gene that appears in the predictor is weighted by how well that gene discriminates between the two classes. However, the selection and weighting of genes are different for the two methods.

Once the value of the compound covariate is computed for each specimen in the training set, a classification threshold is calculated. We use as a threshold $C_t = (C_1 +$

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. A new specimen is predicted to be of class 1 if its compound covariate is closer to C_1 and to be of class 2 if its value is closer to C_2. This prediction rule is equivalent to performing a one-dimensional linear discriminant analysis on the compound covariate with equal prior probabilities for the two classes (13).

Application to gene expression data from breast cancer patients. The compound covariate predictor was previously applied in a study of gene expression in hereditary breast cancer (6). Fluorescence-intensity ratios were calculated and gene-expression profiles generated for 22 primary human breast tumors (7 BRCA1-mutation-positive, 8 BRCA2-mutation-positive and 7 sporadic patient samples). We used the class prediction paradigm discussed above to perform classification of these 22 breast tumors. Our interest was in determining whether hereditary breast cancers could be classified based solely on their gene-expression profiles. We examined two different groupings of the tumors: the first of these labels the 22 tumors according to BRCA1-mutation status (positive or negative) and the second labels the tumors according to BRCA2-mutation status (again, positive or negative). For each set of class labels we performed classification with the compound covariate predictor using a predefined significance level of $\alpha = 0.0001$ for determining differentially expressed genes. We implemented leave-one-out cross-validation: one specimen was removed from the data set, differentially expressed genes determined for the remaining specimens, compound covariate values computed for each of these, the classification threshold calculated and a prediction made for the label of the left out specimen. After the labels of all specimens were predicted, we performed a permutation analysis to assess the significance of prediction results.

Results of class prediction for both BRCA1 and BRCA2 groupings are shown in Table 2. BRCA1 classification resulted in the correct prediction of 7 out of 7 BRCA1-mutation-positive tumors and 14 out of 15 BRCA1-mutation-negative tumors. BRCA2 classification resulted in the correct class prediction of 5 out of 8 BRCA2-mutation-positive tumors and 13 out of 14 BRCA2-mutation-negative tumors. Moreover, the accuracy of the BRCA1 and BRCA2 classifications were significant when compared to the accuracy attained using randomized data: permutation testing estimated the achieved significance level to be 0.004 for BRCA1 classification and 0.043 for BRCA2 classification. Of great interest, it was subsequently determined that the one patient who was incorrectly classified in the BRCA1 classification (a sporadic cancer predicted to be BRCA1-mutation-positive) was the only patient among those with sporadic breast cancer who had hypermethylation of the BRCA1 promoter region, indicative of BRCA1 inactivation (6).

Discussion

The standard proposed here for the application of class prediction methods is designed to increase the quality of predictions. We establish guidelines for evaluating the appropriateness of class prediction for the research question at hand. If the setting is proper, specifications are made for performing class prediction and assessing the significance of results. The framework does not require an independent set of data for validation, though we strongly encourage authentication on independent data when possible. Indeed, a strength of the paradigm is that it should lead to the formation of prediction rules that maintain their predictive ability when applied to validation sets. Although we make some general comments on selecting a prediction method for use with gene expression data, a determination of which methods consistently perform well has not been made. It may be the case that the performance of methods will be quite dependent on the classes of tumor being studied. At any rate, the framework we present here will serve as a good setting for the direct comparison of prediction methods.

The specific prediction method that we used in this paper, compound covariate prediction, performed well in classifying breast cancers with regard to mutation status. Even so, misclassifications did occur. This is an important point, because the level of accuracy needed for a prediction rule to be clinically useful is likely to be more stringent than what is necessary for determining that gene expression profiles significantly differ between two groups. Hence, we are considering improvements to compound covariate prediction. One option is to more carefully select differentially expressed genes: for example, by explicitly controlling the number of spurious differentially expressed genes included in the predictor. Another option is to carry out prediction on multiple compound covariates, with each compound covariate representing a particular functional group of genes. However, we attempted the latter for breast cancer classification and predictive accuracy declined.

Perhaps achieving a better predictor will require more narrowly defined class labels. For example, it was discovered that one of the sporadic breast cancers we attempted to classify was hypermethylated at the BRCA1 promoter region. This case was misclassified in the BRCA1 analysis but, more importantly, the expression profile of this unusual case may have served as a contaminant in the development of the prediction rule. Although no other cases were misclassified in BRCA1 classification, prediction of new cases may be affected by this contaminant. The more rigorously classes are defined and the more thoroughly specimens are evaluated for inclusion in the training set, the more likely it is that good predictors will result.

Extension of the paradigm we propose to more than two classes is straight-forward. Extension to the prediction of a quantitative phenotype such as survival time is also possible, but a cross-validated prediction accuracy must be used instead of a cross-validated misclassification rate.

References

1. Brazma, A. & Vilo, J. (2000) *FEBS Lett.* **480,**17-24.
2. Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M. L., Downing, J. R., Caligiuri, M. A., *et al.* (1999) *Science* **286,** 531-537.
3. Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S., Ares, M., Jr. & Haussler, D. (2000) *Proc. Natl. Acad. Sci. USA* **97,** 262-267.
4. Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M. & Haussler, D. (2000) *Bioinformatics* **16,** 906-914.
5. Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. & Yakhini, Z. (2000) *J. Comput. Biol.* **7,** 559-583.
6. Hedenfalk, I., Duggan, D., Chen, Y., Radmacher, M., Bittner, M., Simon, R., Meltzer, P., Gusterson, B., Esteller, M., Raffeld, M., *et al*. (2001) *N. Engl. J. Med.* **344,** 539-548.
7. Hills, M. (1966) *J. R. Stat. Soc. [B]* **28,** 1-31.
8. Lachenbruch, P. A. & Mickey, M. R. (1968) *Technometrics* **10,** 1-11.
9. Efron, B. (1983) *J. Am. Stat. Assoc.* **78,** 316-331.
10. Lehmann, E. L. & Stein, C. (1949) *Ann. Math. Stat.* **20,** 28-45.
11. Efron, B. & Tibshirani, R. J. (1998) *An introduction to the bootstrap* (Chapman & Hall/CRC, New York), pp. 202-210.
12. Tukey, J. W. (1993) *Controlled Clin. Trials* **14,** 266-285.
13. Morrison, D. F. (1967) *Multivariate statistical methods* (McGraw-Hill, New York), pp. 130-133.

Table 1. Percentage of simulated data sets*

with *m* or fewer misclassifications.

		Leave-one-out
m	Non-cross-validated	Cross-validated
0	99.85	0.60
1	100.00	2.70
2	100.00	6.20
3	100.00	11.20
4	100.00	16.90
5	100.00	24.25
6	100.00	34.00
7	100.00	42.55
8	100.00	53.85
9	100.00	63.60
10	100.00	74.55
11	100.00	83.50
12	100.00	91.15
13	100.00	96.85
14	100.00	100.00

* 2000 data sets were simulated as described in text.

Table 2. Classification of hereditary breast cancers with the compound covariate predictor.

Class labels	m = number of cross-validated misclassifications	Proportion of random permutations with m or fewer misclassifications[*]
BRCA1$^+$, BRCA1$^-$	1 (0 BRCA1$^+$, 1 BRCA1$^-$)	0.004[†]
BRCA2$^+$, BRCA2$^-$	4 (3 BRCA2$^+$, 1 BRCA2$^-$)	0.043[†]

[*] Serves as an estimate of the achieved significance level (see text for details).

[†] These values are slightly different than those appearing in (6) because here, twice as many permutations (i.e., 2000) were performed for each classification.