Pitfalls in the Use of DNA Microarray Data for Diagnostic and Prognostic Classification

Richard Simon, Michael D. Radmacher, Kevin Dobbin, Lisa M. McShane

DNA microarrays have made it possible to estimate the level of expression of thousands of genes for a sample of cells. Although biomedical investigators have been quick to adopt this powerful new research tool, accurate analysis and interpretation of the data have provided unique challenges. Indeed, many investigators are not experienced in the analytical steps needed to convert tens of thousands of noisy data points into reliable and interpretable biologic information. Although some investigators recognize the importance of collaborating with experienced biostatisticians to analyze microarray data, the number and availability of experienced biostatisticians is inadequate. Consequently, investigators are using available software to analyze their data, many seemingly without knowledge of potential pitfalls. Because of serious problems associated with the analysis and reporting of some DNA microarray studies, there is great interest in guidance on valid and effective methods for analysis of DNA microarray data.

The design and analysis strategy for a DNA microarray experiment should be determined in light of the overall objectives of the study. Because DNA microarrays are used for a wide variety of objectives, it is not feasible to address the entire range of design and analysis issues in this commentary. Here, we address statistical issues that arise from the use of DNA microarrays for an important group of objectives that has been called "class prediction" *(1)*. Class prediction includes derivation of predictors of prognosis, response to therapy, or any phenotype or genotype defined independently of the gene expression profile.

Experimental Objectives Drive Design and Analysis

Good DNA microarray experiments, although not based on gene-specific mechanistic hypotheses, should be planned and conducted with clear objectives. Three commonly encountered types of study objectives are "class comparison," "class prediction," and "class discovery" *(1)*.

Class comparison is the comparison of gene expression in different groups of specimens. The major characteristic of class comparison studies is that the classes being compared are defined independently of the expression profiles. The specific objectives of such a study are to determine whether the expression profiles are different between the classes and, if so, to identify the differentially expressed genes. One example of a class comparison study is the comparison of gene expression profiles of stage I breast cancer patients who are long-term survivors with the gene expression profiles of those who have recurrent disease. Another example is the comparison between gene expression profiles in breast cancer patients with and without germline BRCA1 mutations *(2)*.

Class prediction studies are similar to class comparison studies in that the classes are predefined. In class prediction studies, however, the emphasis is on developing a gene expression-based multivariate function (referred to as the predictor) that accurately predicts the class membership of a new sample on the basis of the expression levels of key genes. Such predictors can be used for many types of clinical management decisions, including risk assessment, diagnostic testing, prognostic stratification, and treatment selection. Many studies include both class comparison and class prediction objectives.

Class discovery is fundamentally different from class comparison or class prediction in that no classes are predefined. Usually the purpose of class discovery in cancer studies is to determine whether discrete subsets of a disease entity can be defined on the basis of gene expression profiles. This purpose is different from determining whether the gene expression profiles correlate with some already known diagnostic classification. Examples of class discovery are the studies by Bittner et al. *(3)* that examined gene expression profiles for advanced melanomas and by Alizadeh et al. *(4)* that examined the gene expression profiles of patients with diffuse large B-cell lymphoma. Often the purpose of class discovery is to identify clues regarding the heterogeneity of disease pathogenesis.

Limitations of Cluster Analysis for Class Prediction

One of the most common errors in the analysis of DNA microarray data is the use of cluster analysis and simple fold change statistics for problems of class comparison and class prediction. Although cluster analysis is appropriate for class discovery, it is often not effective for class comparison or class prediction. Cluster analysis refers to an extensive set of methods for partitioning samples into groups on the basis of the similarities and differences (referred to as distances) among their gene expression profiles. Because there are many ways of measuring distances among gene expression profiles involving thousands of genes and because there are many algorithms for partitioning, cluster analysis is a very subjective analysis strategy.

Cluster analysis is considered an unsupervised method of analysis because no information about sample grouping is used. The distance measures are generally computed with regard to the complete set of genes represented on the array that are measured with sufficiently high signals, or with regard to all the genes that

Affiliations of authors: R. Simon, K. Dobbin, L. M. McShane, Biometric Research Branch, National Cancer Institute, National Institutes of Health, Bethesda, MD; M. D. Radmacher, Departments of Biology and Mathematics, Kenyon College, Gambier, OH.

Correspondence to: Richard Simon, D.Sc., National Cancer Institute, 9000 Rockville Pike, MSC 7434, Bethesda, MD 20892-7434 (e-mail: rsimon@nih.gov).

See "Note" following "References."

show meaningful variation across the sample set. Because relatively few genes may distinguish any particular class, the distances used in cluster analysis will often not reflect the influence of these relevant genes. This feature accounts for the poor results often obtained in attempting to use cluster analysis for class prediction studies.

Cluster analysis also does not provide statistically valid quantitative information about which genes are differentially expressed between classes. Investigators often use simple average fold change measures or visual inspection of a cluster image display to identify differentially expressed genes. However, average fold change indices do not account for variability in gene expression across samples within the same class; some twofold average effects represent statistically significant differences and some do not. Neither fold change indices nor visual inspection of cluster image displays enable the investigator to deal with multiple comparison issues in a statistically valid manner. For example, in examining expression levels of thousands of randomly varying genes, there may be many genes that spuriously appear to be differentially expressed between two classes on the basis of visual inspection or fold change thresholds.

Class Prediction Using Supervised Methods

For class prediction studies, it is more appropriate to use a supervised method (i.e., one that makes distinctions among the specimens on the basis of predefined class label information) than an unsupervised method, such as cluster analysis. Supervised class prediction is usually based on the assumption that a collection of differentially expressed genes is associated with class distinction.

The first step toward constructing the class predictor (sometimes called the classifier) is to select the subset of informative genes. The second step is often to assign weights related to the individual predictive strengths of these informative genes. Predictors based on linear combinations of the weighted intensity measurements of the informative genes have been proposed *(1,5)*. One alternative method is to use a dimension reduction technique such as principal components analysis or partial least squares on the informative genes and to base the prediction on the resulting factors *(6–8)*. Many other methods for defining a multivariate predictor have been described *(9,10)*. The final step in constructing the classifier is to define the prediction rule. For example, in a two-group classification where a single predictor is computed, the classification rule may simply be a threshold value; a specimen is classified as being in one group if the derived predictor value is less than the threshold and classified as being in the other group if the derived predictor value is more than the threshold.

One major limitation of supervised methods is overfitting the predictor. Overfitting means that the number of parameters of the model is too large relative to the number of cases or specimens available. Because the model parameters are optimized for the data, the model will fit the original data but may predict poorly for independent data. This happens because the model fits random variations within the original data that do not represent true relationships that hold for independent data. Consequently, it is essential to obtain an unbiased estimate of the true error rate of the predictor (i.e., the probability of incorrectly classifying a randomly selected future case).

Methods for obtaining unbiased estimates of a predictor's error rate include leave-one-out cross-validation or application of the prediction rule developed from a supervised analysis of one dataset to an independent dataset. By using these techniques, it is possible not only to evaluate overfitting the predictor, but also to compare various prediction methods and assess which ones are less prone to overfitting. The appropriate use of leave-one-out cross-validation and validation of independent datasets is discussed in the next two sections.

Cross-Validation of Prediction Accuracy

The performance of a class prediction rule is best assessed by applying the rule created on one set of data (the training set) to an independent set of data (the validation set). Because most clinical research laboratories have access to only a limited number of tumor samples, withholding a substantial proportion of the samples from the training set for the sake of creating a validation set may considerably reduce the performance of the prediction rule. Cross-validation procedures use the data more efficiently. A small number of specimens are withheld, and most of the specimens are used to build a predictor. The predictor is used to predict class membership for the withheld specimens. This process is iterated, leaving out a new set of specimens at each step, until all specimens have been classified. In leave-one-out cross-validation, for example, each specimen is excluded from the training set one at a time and then classified on the basis of the predictor built from the data for all of the other specimens. The leave-one-out cross-validation procedure provides a nearly unbiased estimate of the true error rate of the classification procedure. The estimated error rate applies to the procedure used to build the classifier rather than to the specific prediction model based on all the data, because there is a different classifier for each leave-one-out training set *(11,12)*. Other cross-validation methods omit more than one specimen at a time *(13)* and also produce nearly unbiased estimates.

In the previous section, three common components of class prediction methods were listed: 1) selection of informative genes, 2) computation of weights for selected informative genes, and 3) creation of a prediction rule. It is important that all three steps undergo the cross-validation procedure. Failure to cross-validate all steps may lead to substantial bias in the estimated error rate.

We performed a simulation to examine the bias in estimated error rates for a class prediction study with various levels of cross-validation (*see* supplemental information at http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue1/index.shtml and at http://linus.nci.nih.gov/~brb for a full description of the simulation). We considered two types of leave-one-out cross-validation: one with removal of the left-out specimen before selection of differentially expressed genes and one with removal of the left-out specimen after gene selection but before computation of gene weights and application of the prediction rule. We also computed the resubstitution estimate of the error rate (this estimate results from building the predictor on the full dataset and then reapplying it to each specimen for classification purposes). In each simulated dataset, 20 expression profiles of 6000 genes were randomly generated from the same distribution. Ten profiles were arbitrarily assigned to class 1 and the other 10 profiles to class 2, creating an artificial separation of the profiles into two classes. Because no true underlying difference existed between the two classes, the class prediction should perform no better than a random guess, with estimated error rates for simulated datasets centering around 0.5 (i.e., 10 misclassifications of 20).

The observed number of misclassifications resulting from each level of cross-validation for 2000 simulated datasets is shown in Fig. 1. The resubstitution estimate of the error is biased for small datasets *(11,12)*; our simulation confirms this, with 98.2% of the simulated datasets resulting in zero misclassifications, even though no true underlying difference existed between the two groups. Moreover, the maximum number of misclassified specimens obtained using the resubstitution method was one. Cross-validating the prediction rule after selecting differentially expressed genes from the full dataset does little to correct the bias, with 90.2% of simulated datasets still resulting in zero misclassifications. However, when gene selection is also subjected to cross-validation, we observed results in line with our expectation: the median number of misclassified profiles was 11, although the range was large (0–20).

The simulation results underscore the importance of cross-validating all steps of predictor construction in estimating a true error rate. A recently published study *(14)* also illustrates the point. van't Veer et al. *(14)* predicted clinical outcome of patients with axillary lymph node-negative breast cancer (metastatic disease within 5 years versus disease-free at 5 years) from gene expression profiles, first by using an incomplete cross-validation method and then by using a fully cross-validated method. Their incomplete cross-validation method did not include reselection of the differentially expressed genes. The investigators controlled for the number of misclassified recurrent patients (i.e., the sensitivity of the test) in both situations. To illustrate the importance of proper cross-validating, we focus attention on the difference in estimated error rates for the disease-free patients. The incomplete cross-validation method and the fully cross-validated leave-one-out method result in estimated error rates of 27% (12 of 44) and 41% (18 of 44), respectively. The incomplete method results in a seriously biased underestimate of the error rate, probably largely from overfitting the predictor to the specific dataset.

Although van't Veer et al. *(14)* report both the partially and fully cross-validated estimates of the error rate, it is the smaller and invalid partially cross-validated estimate that has received more attention *(15)*. When cross-validation methods are improperly performed, i.e., without repeating all steps of gene selection and predictor construction within each stage of the cross-validation, the results can be almost as biased as if cross-validation had not been used. Unfortunately, this error is common. In a recent study *(16)* where decision trees were built from gene expression data to classify specimens as normal colon or colon cancer, the authors used a procedure that cross-validated only the steps that occurred after selection of the informative genes. The full dataset was used to identify the informative genes.

Validation on Independent Data

A class predictor that results in a small, properly cross-validated error rate for a collection of tumor specimens is a potentially important finding but one that still requires further validation. This is especially true for class prediction studies based on fewer than 50 tumor specimens. The relatively small sample sizes necessitate validation of the predictor on independently collected data for at least two reasons. First, although cross-validated error estimates are nearly unbiased, they have a large variance for small sample sizes *(17)*. For example, a cross-validated error rate of 0.10 derived from a set of 20 tumors may have a large associated standard error and does not guarantee a true error rate of 0.10 for the predictor. The standard error can be reduced somewhat by using more complex versions of cross-validation *(17)*. Second, the tumors used to build the predictor in the original study may not accurately reflect all characteristics of the underlying populations of interest; the predictor may ignore important properties of the larger population or heavily weight peculiarities of the training set.

An independent validation dataset should ideally be large enough to demonstrate statistically that predictions are accurate.

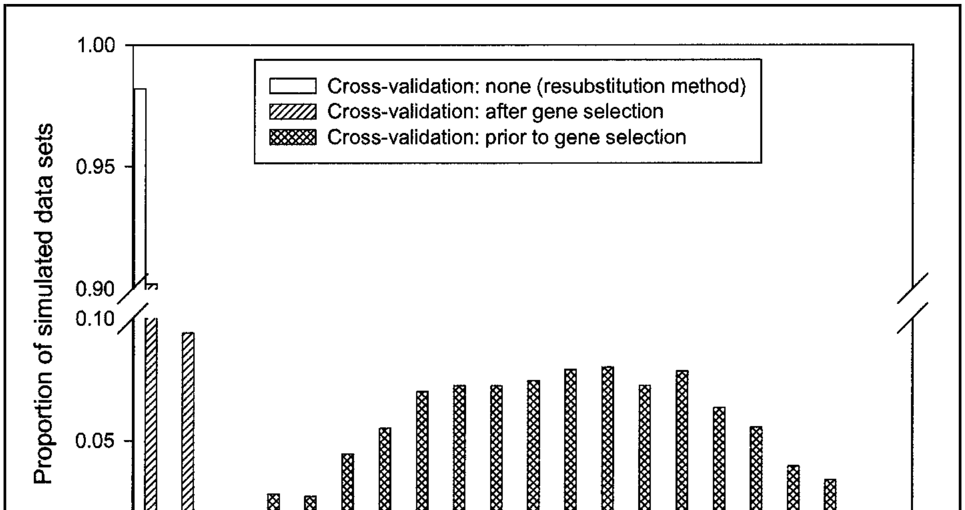

Fig. 1. Effect of various levels of cross-validation on the estimated error rate of a predictor derived from 2000 simulated datasets. Class labels were arbitrarily assigned to the specimens within each dataset, so poor classification accuracy is expected. Class prediction was performed on each dataset as described in the supplemental information (http://jncicancerspectrum.oupjournals.org/jnci/content/vol95/issue1/index.shtml and http://linus.nci.nih.gov/~brb), varying the level of leave-one-out cross-validation used in the prediction. **Vertical bars** indicate the proportion of simulated datasets (of 2000) resulting in a given number of misclassifications for a specified cross-validation strategy.

Tumor class prediction results are often reported for very small validation sets. For example, from the gene expression profiles of 23 medulloblastoma specimens, MacDonald et al. *(18)* built a predictor that distinguished between metastatic and nonmetastatic specimens, then validated the predictor on five other specimens (although a prediction was made for only four of the five validation specimens). All four of the validation specimens for which a prediction was made were correctly classified. Although this is a promising result, it provides little information about the accuracy of the predictor: An exact one-sided 95% binomial confidence interval of the true error rate, given zero misclassifications of four specimens in the validation set, ranged from 0 to 0.53. On the basis of this validation set, we cannot confidently state that the predictor performs better than a guess, even though all specimens were correctly classified. Clearly, a larger validation set was needed.

It is also important that performance of the predictor be validated on all the classes for which it was created, with enough specimens from each class. The validation set of MacDonald et al. *(18)* included only nonmetastatic specimens, so no insight is gained from the validation set concerning the predictor's performance on metastatic specimens, which are more difficult to classify. However, returning to the study of clinical outcome in breast cancer by van't Veer et al. *(14),* a validation set of nineteen specimens was examined, which contained seven specimens from patients who were disease-free at 5 years. The true error rate of disease-free specimens is the more interesting rate in this situation, because the error rate for the metastatic specimens was controlled to be rather small. Two of the seven disease-free specimens in the validation set were misclassified, resulting in an exact two-sided 95% binomial confidence interval for the true error rate of disease-free specimens ranging from 0.04 to 0.71. A more balanced number of recurrent and disease-free cases in the validation set would have provided more information about the predictor's accuracy.

Reporting the Error Rate

We recommend reporting only properly cross-validated error rates or error rates derived from sufficiently large independent validation sets. Some predictors allow a specimen to remain unclassified if the specimen cannot confidently be assigned to any of the examined classes. We suggest that reported error rates account for these unclassified specimens. For example, MacDonald et al. *(18)* developed a cross-validated predictor on the basis of the weighted voting method of Golub et al. *(1).* A prediction strength index was assigned to each sample, with numbers close to (+1) indicating a confident prediction of belonging to the nonmetastatic group, and numbers close to (–1) indicating a confident prediction of belonging to the metastatic group. If the absolute value of the prediction strength index for a specimen did not exceed a threshold of 0.23, then the specimen was assigned to an uncertain group. The authors emphasized that their predictor had an accuracy of 72% *(18).* However, in this calculation they did not count samples that their methodology classified as uncertain. Although "uncertain" may be a clinically important category, it may also be seen as a failure of the classification procedure. For example, a predictor that classifies one sample correctly and calls the rest uncertain has an accuracy of 100% by disregarding unclassified specimens but, at the same time, is of little practical value. Simply ignoring the unclassified specimens gives an overly optimistic impression of the predictor. The percentage of specimens that were correctly classified in the MacDonald et al. *(18)* study was 57%, and the percentage of specimens that were misclassified was 22%; these seem to be more pertinent statistics.

Claims for New Class Prediction Methods

A wide variety of class prediction algorithms, including complex new algorithms, are being applied to DNA microarray data. However, in settings where the number of candidate predictors is orders of magnitude greater than the number of cases, complex methods with many parameters often do not perform well when properly evaluated. Comparisons of class predictors varying in degree of complexity conducted by Dudoit et al. *(10)* demonstrate this finding specifically for DNA microarray data. Simple methods such as diagonal linear discriminant analysis and nearest neighbor classification *(18),* the weighted voting method *(1),* and the compound covariate predictor *(2,5)* have been very effective in cancer studies with small numbers of cases. Some authors have made strong claims about the value of a new prediction algorithm but with no comparison to other algorithms [e.g., *see (6)*]. Moreover, some classes are very easy to distinguish on the basis of gene expression profiles, but the effectiveness of an algorithm cannot be evaluated without comparing it to other algorithms using the same dataset.

Conclusion

Many studies profiling gene expression in human cancers have been completed and are in progress. Some studies *(19–22)* attempt to build predictors of patient prognosis and response to therapy by using gene expression profiles. Because it is likely that gene expression profiles will provide information that will affect clinical decision making, such profiling studies must be performed with statistical rigor and be reported clearly and with unbiased statistics. We recommend that supervised methods rather than cluster analyses be used for class prediction and class comparison studies. Cluster analyses are less powerful than supervised methods for distinguishing predefined classes, and they do not provide valid statistical identification of differentially expressed genes. Biased resubstitution or only partially cross-validated estimates should either not be reported or should be clearly represented as unreliable indicators of prediction accuracy. If cross-validation is used to estimate prediction accuracy, then the entire model-building process, including the selection of informative genes, should be repeated in each cross-validation training set. If a separate dataset is used for validation, it should be sufficiently large to provide meaningful confidence intervals for prediction accuracy. We recommend that investigators include all test cases in their reported estimates of prediction accuracy and not exclude those that do not give a clear-cut prediction. Finally, we urge investigators not to make strong claims about the value of new prediction algorithms without comparing them to more standard prediction methods.

References

(1) 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;286:531–7.

(2) Hedenfalk I, Duggan D, Chen Y, Radmacher M, Bittner M, Simon R, et al. Gene expression profiles of hereditary breast cancer. N Engl J Med 2001; 344:539–48.

(3) Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, Hendrix M, et al. Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 2000;406:536–40.

(4) Alizadeh AA, Eisen MB, Davis RE, Ma C, Lossos IS, Rosenwald A, et al. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature 2000;403:503–11.

(5) Radmacher MD, McShane LM, Simon R. A paradigm for class prediction using gene expression profiles. J Comput Biol 2002;9:505–12.

(6) Khan J, Wei JS, Ringnér M, Saal LH, Ladanyi M, Westermann F, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. Nat Med 2001;7:673–9.

(7) West M, Blanchette C, Dressman H, Huang E, Ishida S, Spang R, et al. Predicting the clinical status of human breast cancer by using gene expression profiles. Proc Natl Acad Sci U S A 2001;98:11462–7.

(8) Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. Bioinformatics 2002;18:39–50.

(9) Ripley BD. Pattern recognition and neural networks. Cambridge (U.K.): Cambridge University Press; 1996.

(10) Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77–87.

(11) Hills M. Allocation rules and their error rates. J R Stat Soc B 1966;28:1–31.

(12) Lachenbruch PA, Mickey MR. Estimation of error rates in discriminant analysis. Technometrics 1968;10:1–11.

(13) Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. Belmont (CA): Wadsworth International Group; 1984.

(14) van't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, et al. Gene expression profiling predicts clinical outcome of breast cancer. Nature 2002;415:530–6.

(15) Caldas C, Aparicio AJ. The molecular outlook. Nature 2002;415:484–5.

(16) Zhang H, Yu CY, Singer B, Xiong M. Recursive partitioning for tumor classification with gene expression microarray data. Proc Natl Acad Sci U S A 2001;98:6730–5.

(17) Efron B, Tibshirani R. Improvements on cross-validation: the .632+ bootstrap method. J Am Stat Assoc 1997;92:548–60.

(18) MacDonald TJ, Brown KM, LaFleur B, Peterson K, Lawlor C, Chen Y, et al. Expression profiling of medulloblastoma: PDGFRA and the RAS/MAPK pathway as therapeutic targets for metastatic disease. Nat Genet 2001;29:143–52.

(19) Garber ME, Troyanskaya OG, Schluens K, Petersen S, Thaesler Z, Pacyna-Gengelbach M, et al. Diversity of gene expression in adenocarcinoma of the lung. Proc Natl Acad Sci U S A 2001;98:13784–9.

(20) Bhattacharjee A, Richards WG, Staunton J, Li C, Monti S, Vasa P, et al. Classification of human lung carcinomas by mRNA expression profiling reveals distinct adenocarcinoma sub-classes. Proc Natl Acad Sci U S A 2001;98:13790–5.

(21) Shipp MA, Ross KN, Tamayo P, Weng AP, Kutok JL, Aguiar RC, et al. Diffuse large B-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nat Med 2002;8:68–74.

(22) Rosenwald A, Wright G, Chan WC, Connors JM, Campo E, Fisher RI, et al. The use of molecular profiling to predict survival after chemotherapy for diffuse large-B-cell lymphoma. New Engl J Med 2002;346:1937–47.

Note

Manuscript received June 4, 2002; revised October 2, 2002; accepted October 11, 2002.