Comparison of Discrimination Methods for the Classification of Tumors Using Gene Expression Data

Sandrine DUDOIT, Jane FRIDLYAND, and Terence P. SPEED

A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies increasingly used in cancer research. By allowing the monitoring of expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more informative classification. The ability to successfully distinguish between tumor classes (already known or yet to be discovered) using gene expression data is an important aspect of this novel approach to cancer classification. This article compares the performance of different discrimination methods for the classification of tumors based on gene expression data. The methods include nearest-neighbor classifiers, linear discriminant analysis, and classification trees. Recent machine learning approaches, such as bagging and boosting, are also considered. The discrimination methods are applied to datasets from three recently published cancer gene expression studies.

KEY WORDS: Cancer; Discriminant analysis; Microarray experiment; Supervised learning; Tumor classification; Variable selection.

1. INTRODUCTION

A reliable and precise classification of tumors is essential for successful diagnosis and treatment of cancer. Current methods for classifying human malignancies rely on a variety of morphological, clinical, and molecular variables. Despite recent progress, there are still uncertainties in diagnosis. Furthermore, it is likely that the existing classes are heterogeneous and comprise diseases that are molecularly distinct and follow different clinical courses. cDNA microarrays and high-density oligonucleotide chips are novel biotechnologies increasingly used in cancer research (Alon et al. 1999; Golub et al. 1999; Perou et al. 1999; Pollack et al. 1999; Alizadeh et al. 2000; Ross et al. 2000). By allowing the monitoring of expression levels in cells for thousands of genes simultaneously, microarray experiments may lead to a more complete understanding of the molecular variations among tumors and hence to a finer and more reliable classification.

Types of microarray systems include the cDNA microarrays developed in the Brown and Botstein labs at Stanford (DeRisi, Iyer, and Brown 1997; Eisen, Spellman, Brown, and Botstein 1998) and the high-density oligonucleotide chips from the Affymetrix Company (Lockhart et al. 1996); the brief description here focuses on the former. cDNA microarrays consist of thousands of individual DNA sequences printed in a high-density array on a glass microscope slide using a robotic arrayer. The relative abundance of these spotted DNA sequences in two DNA or RNA samples may be assessed by monitoring the differential hybridization of the two samples to the sequences on the array. For mRNA samples, the two samples, or *targets*, are reverse-transcribed into cDNA, labeled using different fluorescent dyes [e.g., a red fluorescent dye (cyanine 5 or Cy5), and a green fluorescent dye (cyanine 3 or Cy3)], then mixed in equal proportions and hybridized with the arrayed DNA sequences, or *probes* (following the definition of probe and target adopted in *The Chipping Forecast* 1999). After this competitive hybridization, the slides are imaged using a scanner and fluorescence measurements are made separately for each dye at each spot on the array. The ratio of the red and green fluorescence intensities for each spot is indicative of the relative abundance of the corresponding DNA probe in the two nucleic acid target samples. (See *The Chipping Forecast* 1999 for a more detailed introduction to the biology and technology of cDNA microarrays and oligonucleotide chips.)

Microarray experiments raise numerous statistical questions in fields as diverse as image analysis, experimental design, cluster and discriminant analysis, and multiple hypothesis testing. Here we focus on the classification of tumors using gene expression data. Three main types of statistical problems are associated with tumor classification: (a) identification of new tumor classes using gene expression profiles, *cluster analysis/unsupervised learning*; (b) classification of malignancies into known classes, *discriminant analysis/supervised learning*; and (c) identification of "marker" genes that characterize the different tumor classes, *variable selection.* Data from these new types of experiments present a "large p, small n" problem; that is, a very large number of variables (genes) relative to the number of observations (tumor samples). The publicly available datasets typically contain expression data on 5,000–10,000 genes for less than 100 tumor samples. Both numbers are expected to grow, the number of genes reaching on the order of 30,000, an estimate for the total number of genes in the human genome.

Sandrine Dudoit is Assistant Professor, Division of Biostatistics, University of California, Berkeley, Berkeley, CA 94720 (E-mail: *sandrine@stat.berkeley.edu*). Jane Fridlyand is a Postdoctoral Scientist at the UCSF Cancer Center, San Francisco, CA 94143 (E-mail: *janef@cc.ucsf.edu*). Terence P. Speed is Professor, Department of Statistics, University of California, Berkeley, Berkeley, CA 94720 (E-mail: *terry@stat.berkeley.edu*). This work was supported in part by an MSRI postdoctoral fellowship, a PMMB Burroughs–Wellcome fellowship, and by National Institutes of Health grant R01GM59506 (TPS). The authors are grateful to Ash Alizadeh, Pat Brown, Mike Eisen, and Doug Ross for providing access to their data. Pablo Tamayo's assistance with the ALL/AML data is greatly appreciated. Finally, the authors thank Leo Breiman, Yoram Gat, David Nelson, Mark van der Laan, and Yee Hwa Yang for many helpful discussions and suggestions, Sam Buttrey for his nearest-neighbor program, and the referees for their comments on an earlier version of this article. Supplementary analyses and figures are available at *http://www.stat.berkeley.edu/users/terry/zarray/html*. Dudoit and Fridlyand contributed equally to this work.

Recent publications on cancer classification using gene expression data have focused mainly on the cluster analysis of both tumor samples and genes and include applications of hierarchical clustering (Alon et al. 1999; Perou et al. 1999; Pollack et al. 1999; Alizadeh et al. 2000; Ross et al. 2000) and partitioning methods such as self-organizing maps (Golub et al. 1999). Alizadeh et al. (2000) used hierarchical clustering to study gene expression in the three most prevalent adult lymphoid malignancies. Ross et al. (2000) also relied on hierarchical clustering to monitor gene expression in the 60 cell lines from the National Cancer Institute's anticancer drug screen. Using acute leukemias as a test case, Golub et al. (1999) explored both the cluster analysis and the discriminant analysis of tumors using gene expression data. For discriminant analysis, or "class prediction," they proposed a "weighted gene voting scheme" that turns out to be a variant of a special case of linear discriminant analysis (Section 2.2). So far, most published articles on tumor classification have applied a single technique to a single gene expression dataset, and it is hard to assess the merits of each method in the absence of a comprehensive comparative study.

This article compares the performance of different discrimination methods for the classification of tumors based on gene expression profiles. These methods include traditional ones, such as nearest-neighbor and linear discriminant analysis, as well as more modern ones, such as classification trees. Recent machine learning approaches, such as bagging and boosting, are also considered. The discrimination methods are applied to three recently published datasets: the leukemia (ALL/AML) dataset of Golub et al. (1999), the lymphoma dataset of Alizadeh et al. (2000), and the 60 cancer cell line (NCI 60) dataset of Ross et al. (2000). The article is organized as follows. Section 2 discusses the discrimination methods considered in the comparision study. The datasets are described in Section 3, along with preliminary data processing steps. The study design for the comparison of the discrimination methods is discussed in Section 4, and the results of the study are presented in Section 5. Finally, our findings are summarized and open questions outlined in Section 6.

2. DISCRIMINATION METHODS

For our purposes, gene expression data on p genes for n tumor mRNA samples may be summarized by an $n \times p$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level of gene (variable) j in mRNA sample (observation) i. The expression levels might be either absolute (e.g., oligonucleotide arrays used to produce the leukemia dataset) or relative to the expression levels of a suitably defined common reference sample (e.g., cDNA microarrays used to produce the lymphoma and NCI 60 datasets). When the mRNA samples belong to known classes (e.g., follicular lymphoma), the data for each observation consist of a *gene expression profile* $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and a class label y_i, that is, of predictor variables $\mathbf{x}_i$ and response y_i. For K tumor classes, the class labels y_i are defined to be integers ranging from 1 to K, and n_k denotes the number of observations belonging to class k. Note that the expression levels x_{ij} are in general highly processed data; the raw data in a microarray experiment consist of image files, and important preprocessing steps include image analysis of these scanned images and normalization. In addition, for the publicly available datasets, the number of tumors n is typically below 100, whereas the number of genes p is several thousands. In the comparison of prediction methods, the number of genes will be substantially reduced by identifying a subset of genes whose expression levels are associated with tumor class (Section 3.4).

A *predictor* or *classifier* for K tumor classes partitions the space $\mathcal{X}$ of gene expression profiles into K disjoint and exhaustive subsets, $A_1, \ldots, A_K$, such that for a sample with expression profile $\mathbf{x} = (x_1, \ldots, x_p) \in A_k$, the predicted class is k. Predictors are built from past experience, that is, from tumor samples known to belong to certain classes. Such observations comprise the *learning set* (LS), $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$. Predictors may then be applied to a *test set* (TS), $\mathcal{T} = \{\mathbf{x}_1, \ldots, \mathbf{x}_{n_T}\}$, to predict the class y_i. expression profile $\mathbf{x}_i$ in the test set for each gene. In the event that the y_i are known for the test set, the predicted and true classes may be compared to estimate the error rate of the predictor. A classifier built from a learning set $\mathcal{L}$ is denoted by $\mathcal{C}(\cdot, \mathcal{L})$; the predicted class for a tumor sample with gene expression profile $\mathbf{x}$ is $\mathcal{C}(\mathbf{x}, \mathcal{L})$. Here, we review briefly a number of well-known discrimination methods. General references on discriminant analysis include works by Mardia, Kent, and Bibby (1979), McLachlan (1992), and Ripley (1996).

2.1 Fisher Linear Discriminant Analysis

First applied by Barnard (1935) at the suggestion of Fisher (1936), *Fisher linear discriminant analysis* (FLDA) is based on finding linear combinations $\mathbf{xa}$ of the gene expression levels $\mathbf{x} = (x_1, \ldots, x_p)$ with large ratios of between-group to within-group sums of squares. (See Mardia et al. 1979 for a detailed presentation of FLDA.) For an $n \times p$ learning set data matrix X, the linear combination $X\mathbf{a}$ of the columns of X has a ratio of between-group to within-group sums of squares given by $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$, where B and W denote the $p \times p$ matrices of between-group and within-group sums of squares and cross-products. The extreme values of $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$ are obtained from the eigenvalues and eigenvectors of $W^{-1}B$. The matrix $W^{-1}B$ has at most $s = \min(K-1, p)$ nonzero eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s$, with corresponding linearly independent eigenvectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_s$. The *discriminant variables* are defined as $u_l = \mathbf{x}\mathbf{v}_l$, $l = 1, \ldots, s$, and in particular, $\mathbf{a} = \mathbf{v}_1$ maximizes $\mathbf{a}'B\mathbf{a}/\mathbf{a}'W\mathbf{a}$.

For a tumor sample with gene expression profile $\mathbf{x} = (x_1, \ldots, x_p)$, let $d_k(\mathbf{x}) = \sum_{l=1}^{s}((\mathbf{x} - \bar{\mathbf{x}}_k)\mathbf{v}_l)^2$ denote its (squared) Euclidean distance, in terms of the discriminant variables, from the $1 \times p$ vector of class k sample means $\bar{\mathbf{x}}_k = (\bar{x}_{k1}, \ldots, \bar{x}_{kp})$ for the learning set $\mathcal{L}$. The predicted class for gene expression profile $\mathbf{x}$ is the class whose mean vector $\bar{\mathbf{x}}_k$, is closest to $\mathbf{x}$ in the space of discriminant variables, that is, $\mathcal{C}(\mathbf{x}, \mathcal{L}) = \arg\min_k d_k(\mathbf{x})$. FLDA is a nonparametric method which also arises in a parametric setting. For $K = 2$ classes, FLDA yields the same classifier as the sample maximum likelihood discriminant rule for multivariate normal class densities with the same covariance matrix (see Section 2.2, case 1 for $K = 2$).

2.2 Maximum Likelihood Discriminant Rules

In a situation where the tumor class conditional densities, $\text{pr}(\mathbf{x}|y=k)$, are known, the *maximum likelihood (ML) discriminant rule* predicts the class of a gene expression profile $\mathbf{x}=(x_1,\ldots,x_p)$ by that which gives the largest likelihood to $\mathbf{x}$, that is, $\mathcal{C}(\mathbf{x})=\arg\max_k \text{pr}(\mathbf{x}|y=k)$. When the class-conditional densities are fully known, a learning set is not needed, and the classifier is simply $\mathcal{C}(\mathbf{x})$. In practice, however, even if the parametric form of the class conditional densities is known, the parameters must be estimated from a learning set. Using parameter estimates in place of the unknown parameters yields the *sample ML discriminant rule*.

For multivariate normal class densities, that is, for $\mathbf{x}|y=k\sim N(\mu_k,\Sigma_k)$, the ML discriminant rule is $\mathcal{C}(\mathbf{x})=\arg\min_k\{(\mathbf{x}-\mu_k)\Sigma_k^{-1}(\mathbf{x}-\mu_k)'+\log|\Sigma_k|\}$. In general, this is a quadratic discriminant rule. Interesting special cases include:

1. *Linear discriminant analysis (LDA)*. When the class densities have the same covariance matrix, $\Sigma_k=\Sigma$, the discriminant rule is based on the square of the Mahalanobis distance and is linear in $\mathbf{x}$, and given by $\mathcal{C}(\mathbf{x})=\arg\min_k(\mathbf{x}-\mu_k)\Sigma^{-1}(\mathbf{x}-\mu_k)'$.
2. *Diagonal quadratic discriminant analysis (DQDA)*. When the class densities have diagonal covariance matrices, $\Delta_k=\text{diag}(\sigma_{k1}^2,\ldots,\sigma_{kp}^2)$, the discriminant rule is given by additive quadratic contributions from each gene, that is, $\mathcal{C}(\mathbf{x})=\arg\min_k\sum_{j=1}^p\{(x_j-\mu_{kj})^2/\sigma_{kj}^2+\log\sigma_{kj}^2\}$.
3. *Diagonal linear discriminant analysis (DLDA)*. In this simplest case, when the class densities have the same diagonal covariance matrix $\Delta=\text{diag}(\sigma_1^2,\ldots,\sigma_p^2)$, the discriminant rule is linear and given by $\mathcal{C}(\mathbf{x})=\arg\min_k\sum_{j=1}^p(x_j-\mu_{kj})^2/\sigma_j^2$.

For the corresponding sample ML discriminant rules, the population mean vectors and covariance matrices are estimated from a learning set $\mathcal{L}$ by the sample mean vectors and covariance matrices, $\hat{\mu}_k=\bar{\mathbf{x}}_k$ and $\widehat{\Sigma}_k=S_k$. For the constant covariance matrix case, the pooled estimate of the common covariance matrix $\widehat{\Sigma}=\sum_k(n_k-1)S_k/(n-K)$ is used. In one of the first applications of a discrimination method to gene expression data, Golub et al. (1999) proposed a "weighted gene voting scheme" for binary classification. This method turns out to be a variant of the sample ML rule corresponding to special case 3. For two classes $k=1$ and 2, the sample ML rule assigns a tumor with gene expression profile $\mathbf{x}=(x_1,\ldots,x_p)$ to class 1 if and only if

$$\sum_{j=1}^p\frac{(x_j-\bar{x}_{2j})^2}{\hat{\sigma}_j^2}\geq\sum_{j=1}^p\frac{(x_j-\bar{x}_{1j})^2}{\hat{\sigma}_j^2},$$

that is,

$$\sum_{j=1}^p\frac{(\bar{x}_{1j}-\bar{x}_{2j})}{\hat{\sigma}_j^2}\left(x_j-\frac{(\bar{x}_{1j}+\bar{x}_{2j})}{2}\right)\geq 0.$$

The discriminant function can be rewritten as $\sum_j v_j$, where $v_j=a_j(x_j-b_j)$, $a_j=(\bar{x}_{1j}-\bar{x}_{2j})/\hat{\sigma}_j^2$, and $b_j=(\bar{x}_{1j}+\bar{x}_{2j})/2$. This is almost the same function as used by Golub et al. except for a_j, which they define as $a_j=(\bar{x}_{1j}-\bar{x}_{2j})/(\hat{\sigma}_{1j}+\hat{\sigma}_{2j})$. The quantity $\hat{\sigma}_{1j}+\hat{\sigma}_{2j}$ is an unusual estimate of the standard error of a difference and having standard deviations instead of variances in the denominator of a_j produces the wrong units for the discriminant function. For each prediction made by the classifier, Golub et al. also define a prediction strength (PS) that indicates the "margin of victory": $\text{PS}=(\max(V_1,V_2)-\min(V_1,V_2))/(\max(V_1,V_2)+\min(V_1,V_2))$, where $V_1=\sum_j\max(v_j,0)$ and $V_2=\sum_j\max(-v_j,0)$. Golub et al. chose a conservative prediction strength threshold of .3, below which no predictions are made.

2.3 Nearest-Neighbor Classifiers

Nearest-neighbor (NN) methods are based on a *distance function* for pairs of tumor mRNA samples, such as the Euclidean distance or one minus the correlation of their gene expression profiles. The *k nearest-neighbor* rule, due to Fix and Hodges (1951), proceeds as follows to classify test set observations on the basis of the learning set. For each tumor sample in the test set (a) find the k closest tumor samples in the learning set, and (b) predict the class by majority vote; that is, choose the class that is most common among those k neighbors.

The number of neighbors k is chosen by *cross-validation*; that is, by running the NN classifier on the learning set only. Each tumor sample in the learning set is treated in turn as if it were in the test set; its distance to all of the other learning set tumor samples (except itself) is computed, and it is classified by the NN rule. The classification for each learning set observation is then compared to the truth to produce the cross-validation error rate. This is done for a number of k's (here $k\in\{1,3,5,\ldots,21\}$), and the k for which the cross-validation error rate is smallest is retained for use on the test set.

2.4 Classification Trees

Binary tree structured classifiers are constructed by repeated splits of subsets (nodes) of the space of gene expression profiles $\mathcal{X}$ into two descendant subsets, starting with $\mathcal{X}$ itself. Each terminal subset is assigned a class label, and the resulting partition of $\mathcal{X}$ corresponds to the classifier. There are three main aspects to tree construction: (a) selection of the splits, so that the data in each of the descendant subsets are "purer" than the data in the parent subset; (b) the decision to declare a node terminal, which is done using cross-validation to "prune" the tree; and (c) the assignment of each terminal node to a class. Different tree classifiers use different approaches to deal with these three issues. Here we use the *classification and regression trees* (CART) method described in detail by Breiman, Friedman, Olshen, and Stone (1984) and implemented in the CART software version 1.310 (also in S-PLUS and R function `tree()`). In our comparison study, single pruned trees are grown using 10-fold cross-validation for estimating the classification error.

2.5 Aggregating Classifiers

Breiman (1996, 1998a) found that gains in accuracy could be obtained by *aggregating predictors* built from perturbed versions of the learning set. The key to improved accuracy is the possible instability of a prediction method (i.e., whether

small changes in the learning set result in large changes in the predictor), and unstable procedures tend to benefit the most from aggregation. Classification trees tend to be unstable, whereas, for example, NN classifiers tend to be stable. Thus in this study, only the CART predictors are aggregated. The trees used for aggregation are maximal "exploratory" trees, in the sense that they are grown until each terminal node contains observations from only a single class (Breiman 1996, 1998a).

More precisely, let $\mathcal{C}(\cdot, \mathcal{L}_b)$ denote the classifier built from the bth perturbed learning set $\mathcal{L}_b$ and let w_b denote the weight given to predictions made by this classifier. The predicted class for a tumor sample with gene expression profile $\mathbf{x}$ is obtained by *weighted voting* and given by $\arg\max_k \sum_b w_b I(\mathcal{C}(\mathbf{x}, \mathcal{L}_b) = k)$, where $I(\cdot)$ denotes the indicator function, equaling 1 if the condition in parentheses is true and 0 otherwise. For aggregated classifiers, *prediction votes* (PVs) assessing the strength of a prediction may be defined for each observation. The prediction vote for a gene expression profile $\mathbf{x}$ is defined by $PV(\mathbf{x}) = (\max_k \sum_b w_b I(\mathcal{C}(\mathbf{x}, \mathcal{L}_b) = k))/(\sum_b w_b)$ and $PV \in [0, 1]$. When the perturbed learning sets are given equal weights, that is, $w_b = 1$, the prediction vote is simply the proportion of votes for the "winning" class. Next we describe two main classes of methods for generating perturbed versions of the learning set, bagging and boosting.

2.5.1 Bagging. Nonparametric Bootstrap. In the simplest form of the *b*ootstrap *agg*regat*ing* or *bagging* procedure, perturbed learning sets of the same size as the original learning set are formed by drawing at random with replacement from the learning set, that is, by forming nonparametric bootstrap samples of the learning set. Predictors are built for each perturbed dataset and aggregated by plurality voting ($w_b = 1$). A general problem of the nonparametric bootstrap for small datasets is the discreteness of the sampling space. The method described next gets around this problem by using convex pseudodata.

Convex pseudo-data. Given a learning set $\mathcal{L} = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_{n_L}, y_{n_L})\}$, Breiman (1998b) suggested creating perturbed learning sets based on *convex pseudodata* (CPD). Each perturbed learning set $\mathcal{L}_b$ is generated by repeating the following steps n_L times:

1. Select two instances $(\mathbf{x}, y)$ and $(\mathbf{x}', y')$ at random from the learning set $\mathcal{L}$.
2. Select at random a number v from the interval $[0, d]$, $0 \leq d \leq 1$, and let $u = 1 - v$.
3. The new instance is $(\mathbf{x}'', y'')$, where $y'' = y$ and $\mathbf{x}'' = u\mathbf{x} + v\mathbf{x}'$.

As in standard bagging a classifier is built for each perturbed learning set $\mathcal{L}_b$ and classifiers are aggregated by plurality voting ($w_b = 1$). Note that when the parameter d is 0, CPD reduces to standard bagging, and that the larger the d, the greater the amount of smoothing. In practice, when a test set is not available, d can be chosen by cross-validation.

2.5.2 Boosting. In *boosting*, first proposed by Freund and Schapire (1997), the data are resampled *adaptively* so that the weights in the resampling are increased for those cases most often misclassified. The aggregation of predictors is done by *weighted voting*. Bagging turns out to be a special case of boosting, when the sampling probabilities are uniform at each step and the perturbed predictors are given equal weight in the voting. We have followed an adaptation of Freund and Schapire's *AdaBoost* algorithm, which was described fully in Breiman (1998a) and referred to in his article as *Arc-fs*.

3. DATA AND PREPROCESSING

3.1 Datasets

The different predictors are compared using data from three recently published cancer gene expression studies. For each study, the data have already been processed in several ways, including image analysis of the microarray scanned images, dye normalization, and screening out of genes based on data quality criteria. Because we chose to use publicly available datasets, most of these decisions were beyond our control, and one should bear in mind that different choices could potentially affect the outcome of the comparison (Yang, Dudoit, Luu, and Speed 2001; Yang, Buckley, Dudoit, and Speed 2002).

3.1.1 Lymphoma. This dataset comes from a study of gene expression in the three most prevalent adult lymphoid malignancies: B-cell chronic lymphocytic leukemia (B-CLL), follicular lymphoma (FL), and diffuse large B-cell lymphoma (DLBCL) (see Alizadeh et al. 2000 and http://genome-www.stanford.edu/lymphoma for a detailed description of the experiments). Gene expression levels were measured using a specialized cDNA microarray, the Lymphochip, containing genes that are preferentially expressed in lymphoid cells or that are of known immunologic or oncologic importance. In each hybridization, fluorescent cDNA targets were prepared from a tumor mRNA sample (fluorescent dye Cy5) and a reference mRNA sample derived from a pool of nine different lymphoma cell lines (fluorescent dye Cy3). The cell lines in the common reference pool were chosen to represent diverse expression patterns, so that most spots on the array would exhibit a nonzero signal in the Cy3 channel. This study produced gene expression data for $p = 4{,}682$ genes in $n = 81$ mRNA samples. The mRNA samples consist of 29 cases of B-CLL, 9 cases of FL, and 43 cases of DLBCL. The gene expression data are summarized by an $81 \times 4{,}682$ matrix $X = (x_{ij})$, where x_{ij} denotes the base 2 logarithm of the Cy5/Cy3 background-corrected fluorescence intensity ratio for gene j in lymphoma sample i.

3.1.2 Leukemia. The leukemia dataset was described by Golub et al. (1999) and is available at http://waldo.wi.mit.edu/MPR/data_set_ALL_AML.html. This dataset comes from a study of gene expression in two types of acute leukemias, acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML). Gene expression levels were measured using Affymetrix high-density oligonucleotide arrays containing $p = 6{,}817$ human genes. The data consist of 47 cases of ALL (38 B-cell ALL and 9 T-cell ALL) and 25 cases of AML. Following Golub et al. (Pablo Tamayo, personal communication), three preprocessing steps were applied to the normalized matrix of intensity values available on the website (after pooling the 38 mRNA samples from the learning set and the 34 mRNA samples from the test set): (a) thresholding, floor

of 100 and ceiling of 16,000; (b) filtering, exclusion of genes with $\max/\min \leq 5$ or $(\max - \min) \leq 500$, where max and min refer to the maximum and minimum intensities for a particular gene across the 72 mRNA samples; and (c) base 10 logarithmic transformation. The data were then summarized by a $72 \times 3,571$ matrix $X = (x_{ij})$, where x_{ij} denotes the expression level for gene j in mRNA sample i. Figure 1 displays images of the 72×72 correlation matrix between gene expression profiles for the 72 leukemia mRNA samples.

3.1.3 NCI 60. In this study, cDNA microarrays were used to examine the variation in gene expression among the 60 cell lines from the National Cancer Institute's anticancer drug screen known as NCI 60 (Ross et al. 2000; http://genome-www.stanford.edu/nci60). The 60 cell lines are derived from tumors with different sites of origin: 7 breast, 6 central nervous system (CNS), 7 colon, 6 leukemia, 8 melanoma, 9 non–small–cell lung carcinoma (NSCLC), 6 ovarian, 2 prostate, 8 renal, and 1 unknown (ADR-RES). Gene expression was studied using microarrays with 9,703 spotted cDNA sequences. In each hybridization, fluorescent cDNA targets were prepared from a cell line mRNA sample (fluorescent dye Cy5) and a reference mRNA sample obtained by pooling equal mixtures of mRNA from 12 of the cell lines (fluorescent dye Cy3). To investigate the reproducibility of the entire experimental procedure (e.g., cell culture, mRNA isolation, labeling, hybridization, scanning), a leukemia (K562) cell line and a breast cancer (MCF7) cell line were analyzed by three independent microarray experiments. Ross et al. screened out genes with missing data in more than two arrays. In addition, because of their small class size, the two prostate cell lines and the unknown cell line were excluded from our analysis. The data are summarized by a $61 \times 5,244$ matrix $X = (x_{ij})$, where x_{ij} denotes the base 2 logarithm of the Cy5/Cy3 background-corrected fluorescence intensity ratio for gene j in cell line i.

3.2 Imputation of Missing Data

For the lymphoma and NCI60 datasets, each array contains a number of genes with fluorescence intensity measurements that were flagged by the experimenter and recorded as missing data points. The mean percentage of missing data points per array is 6.6% for the lymphoma dataset and 3.3% for the NCI60 dataset. Some of the discrimination methods discussed in Section 2 are able to handle missing data (e.g., CART), however, others require complete data (e.g., Fisher linear discriminant analysis). Missing data were imputed by a simple k nearest-neighbor algorithm, in which the neighbors are the genes and the distance between neighbors is based on the correlation between their gene expression levels across arrays. For each gene with missing data, (a) compute its correlation with all other $p - 1$ genes, and (b) for each missing array, identify the k nearest genes having data for this array and impute the missing entry by the average of the corresponding entries for the k neighbors. A value of $k = 5$ neighbors was

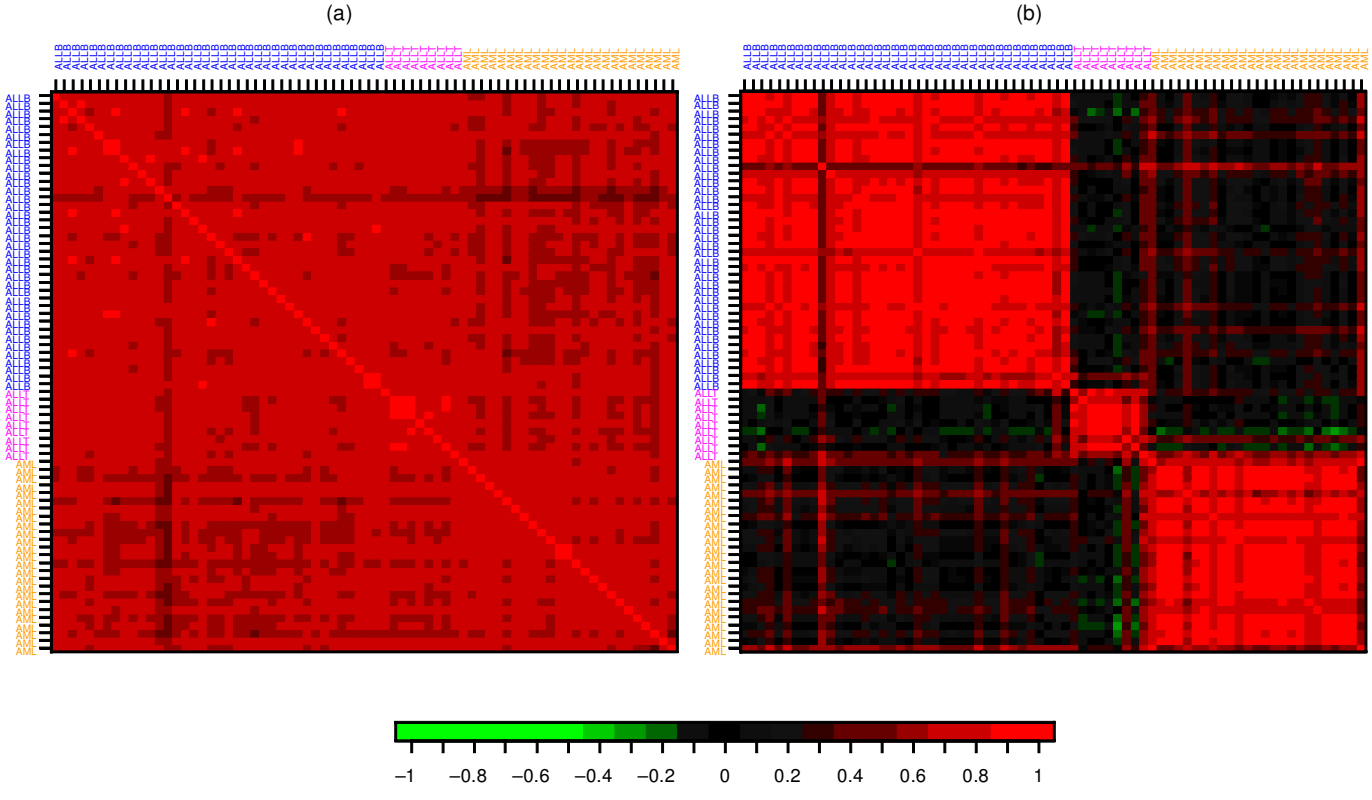

Figure 1. Leukemia Dataset (a) Correlation Matrix. Images of the correlation matrix for the 72 B-cell ALL, T-cell ALL, and AML samples based on expression profiles for $p = 3,571$ genes and (b) for the $p = 40$ genes with the largest BW ratio. Correlations of 0 are represented in black, increasingly positive correlations are represented with reds of increasing intensity, and increasingly negative correlations are represented with greens of increasing intensity. The color bar below the images may be used for calibration purposes. The mRNA samples are ordered by class, first B-cell ALL (blue), then T-cell ALL (magenta), and finally AML (orange).

used for the lymphoma and NCI 60 datasets. For a detailed study of imputation methods in microarray experiments, the reader is referred to the recent work of Troyanskaya et al. (2001), which suggests that a NN approach provides accurate and robust estimates of missing values.

3.3 Standardization

The gene expression data were standardized so that the observations (arrays) have mean 0 and variance 1 across variables (genes). Standardizing the data in this fashion achieves a location and scale *normalization* of the different arrays. In a study of normalization methods, we have found scale adjustment to be desirable in some cases to prevent the expression levels in one particular array from dominating the average expression levels across arrays (Yang et al. 2001). Furthermore, this standardization is consistent with the common practice in microarray experiments of using the correlation between the gene expression profiles of two mRNA samples to measure their similarity (Perou et al. 1999; Alizadeh et al. 2000; Ross et al. 2000).

3.4 Gene Selection

Many genes exhibit near-constant expression levels across tumor samples. We thus performed a preliminary selection of genes based on the ratio of their between-group to within-group sums of squares. For a gene j, this ratio is

$$\mathrm{BW}(j) = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{\cdot j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2},$$

where $\bar{x}_{\cdot j}$ and $\bar{x}_{kj}$ denote the average expression level of gene j across all tumor samples and across samples belonging to class k only. The predictors were built using the p genes with the largest BW ratios. (Section 4 discusses selecting the value of p.)

Note that Golub et al. (1999) used a different method for standardizing the data and for selecting genes. For the sake of completeness, our comparison study includes the weighted gene voting scheme of Golub et al. with a_j calculated using standard deviations instead of variances (Section 2.2) and with the data preprocessed as in Golub et al. (1999). We refer to the resulting predictor as "Golub," and it is of interest to compare its performance with that of DLDA with data preprocessed as in Sections 3.3 and 3.4.

4. STUDY DESIGN

In the absence of genuine test sets, the different predictors are compared based on random divisions of each dataset into a learning set $\mathcal{L}$ and a test set $\mathcal{T}$. There are no widely accepted guidelines for choosing the relative sizes of these artificial learning sets and test sets. A possible choice is to use leave-one-out cross-validation or to leave out a randomly selected 10% of the observations to use as a test set (Breiman 1996). However, in our case, test sets containing only 10% of the data are not large enough to provide adequate discrimination between the classifiers. Because our main purpose is to compare classifiers, not to estimate generalization error, we choose to sacrifice training data and increase test set size to one-third of the data (i.e., a 2:1 scheme).

In the principal comparison, for each learning set/test set (LS/TS) run, the p genes with the largest BW ratio are selected using the learning set. For a comparison involving all predictors, FLDA sets an upper limit on the size of the gene set because of rank issues; the value of p for each dataset was chosen with these constraints in mind: $p = 50$ for the lymphoma dataset, $p = 40$ for the leukemia dataset, and $p = 30$ for the NCI 60 dataset. Next, predictors are constructed using the LS and TS error rates are obtained by applying the predictors to the test set. Aggregated predictors (bagging and boosting) are built from $B = 50$ "pseudo" learning sets, and several values of the parameter d ($d = .05, .1, .25, .5, .75, 1$) were examined for CPD. This entire procedure is repeated $N = 200$ times. Each LS/TS run yields a test set error rate for each predictor; the results are summarized by computing the median error rate for each predictor over the 200 runs (Table 1).

The effect of increasing ($p = 200$) or decreasing ($p = 10$) the number of genes is briefly examined. A "smarter" BW criterion is also applied to the lymphoma data. For $p = 10$ genes, this criterion consists of selecting the five genes with the largest BW ratio (as before) and the five genes with the largest BW ratio when the two largest classes (B-CLL and DLBCL) are pooled. Such a criterion should allow better discrimination of the smaller FL class.

5. RESULTS

5.1 Test Set Error

Table 1 displays the median and upper quartile of the number of misclassified tumor samples for each classifier. For the leukemia dataset, the classifiers are compared based on their ability to distinguish between ALL and AML (two-class problem), and between B-cell ALL, T-cell ALL, and AML (three-class problem). In general, the NN and DLDA predictors had the smallest error rates, whereas FLDA had the highest. With the exception of the NCI 60 dataset, the error rates seem fairly low given the small learning sets.

5.1.1 Nearest-Neighbor Classifiers. The distance function used for the NN classifiers is one minus the correlation between the gene expression profiles of two tumor mRNA samples. The parameter k for the number of neighbors was selected by cross-validation and was usually quite small for each dataset: 1 or 2 for about half of the runs and generally less than 7. Although the small number of neighbors k is an artifact of the small sample sizes, the results also suggest that very good predictions can be obtained from the class of the tumor sample most highly correlated to the sample to be predicted. Indeed, for all three datasets, tumor samples within the same class tended to have strongly and positively correlated gene expression profiles (patches of red along the diagonal of the correlation matrix in Figure 1). This pattern is much more subtle for the correlation matrices based on all genes and, as expected, the NN method benefitted greatly from the initial selection of genes. The pattern is also stronger for the lymphoma and leukemia data than for the NCI 60 data. (See web supplement http://www.stat.berkeley.edu/users/

terry/zarray/Html for figures for the NCI 60 and lymphoma data.)

5.1.2 Fisher Linear Discriminant Analysis. On the opposite end of the performance spectrum is FLDA. The most likely reason for the poor performance of FLDA is that with a limited number of tumor samples and a fairly large number of genes p, the matrices of between-group and within-group sums of squares and cross-products are quite unstable and provide poor estimates of the corresponding population quantities. The performance of FLDA dramatically improves and reaches error rates comparable to DLDA when the number of genes is decreased to $p = 10$, especially when the genes are selected according to the "smarter" BW criterion of Section 4 (see web supplement). Note also that FLDA is a "global" method, that is, it makes use of all of the data for each prediction, and as a result, some tumor samples may not be well represented by the discriminant variables. (There is only one discriminant variable for the two-class leukemia dataset and two discriminant variables for the three-class lymphoma dataset.) In contrast, NN methods are "local."

5.1.3 Maximum Likelihood Discriminant Rules. The simple DLDA rule produced impressively low misclassification rates compared with more sophisticated predictors, such as bagged classification trees. With the exception of the lymphoma dataset, linear classifiers (i.e., DLDA), that assume a common covariance matrix for the different classes, yielded lower error rates than quadratic classifiers (i.e., DQDA), that allow for different class covariance matrices. Thus for the datasets considered here, gains in accuracy were obtained by ignoring correlations between genes. DLDA classifiers are sometimes called "naive Bayes" because they arise in a Bayesian setting, where the predicted tumor class is the one with maximum posterior probability $\text{pr}(y = k|\mathbf{x})$.

5.1.4 Weighted Gene Voting Scheme. For the binary class leukemia dataset, the performance of the variant of DLDA implemented by Golub et al. (1999) was also examined. This method performed similarly to boosting, CPD, and DQDA, but was inferior to NN and especially to DLDA, which had a median error rate of 0. Note that in contrast to the aggregated predictors in bagging and boosting, the "voting" is over variables (here genes) rather than over classifiers. Furthermore, the gene voting scheme as defined by Golub et al. (1999) is applicable to binary classes only; the closest generalization of it to multiple classes is the standard DLDA predictor, which was applied to the other datasets.

5.1.5 Classification Trees. CART-based predictors had performance intermediate between the best classifiers (DLDA, NN) and the worst classifier (FLDA). Aggregated tree predictors were generally more accurate than a single cross-validated tree, with CPD and boosting performing better than standard bagging. Several values of the parameter d, $d = .05, .1, .25, .5, .75$, were tried for the CPD method. For each dataset, the best value turned out to be between .5 and 1, suggesting that the performance of CPD was not very sensitive to the parameter d controlling the degree of smoothing. A value of $d = .75$ was used in Table 1.

Table 1. Test Set Error. Median and Upper Quartiles Over 200 LS/TS Runs, of the Number of Misclassified Tumor Samples for 9 Discrimination Methods Applied to 3 Datasets. For a Given Dataset, the Error Numbers for the Best Predictor are in Bold.

	Leukemia[a]				Lymphoma[b] Three classes		NCI 60[c] Eight classes	
	Two classes		Three classes					
	Median quartile	Upper quartile	Median quartile	Upper quartile	Median quartile	Upper quartile	Median quartile	Upper quartile
Linear and quadratic discriminant analysis								
FLDA[d]	3	4	3	4	6	8	11	11
DLDA[e]	**0**	**1**	1	2	1	1	**7**	**8**
Golub[f]	1	2	–	–	–	–	–	–
DQDA[g]	1	2	1	2	**0**	**1**	9	10
Classification trees								
CV[h]	3	4	1	3	2	3	12	13
Bag[i]	2	2	1	2	2	3	10	11
Boost[j]	1	2	1	2	1	2	9	11
CPD[k]	1	2	1	3	1	2	9	10
Nearest neighbors								
	1	1	**1**	**1**	**0**	**1**	8	10

[a] Leukemia dataset from Golub et al. (1999), test set size $n_{TS} = 24, p = 40$ genes.
[b] Lymphoma dataset from Alizadeh et al. (2000), test set size $n_{TS} = 27, p = 50$ genes.
[c] NCI 60 dataset from Ross et al. (2000), test set size $n_{TS} = 21, p = 30$ genes.
[d] FLDA: Fisher linear discriminant analysis.
[e] DLDA: diagonal linear discriminant analysis.
[f] Golub: weighted gene voting scheme of Golub et al. (1999).
[g] DQDA: diagonal quadratic discriminant analysis.
[h] CV:single CART tree with pruning by 10-fold cross-validation.
[i] Bag:$B = 50$ bagged exploratory trees.
[j] Boost: $B = 50$ boosted exploratory trees.
[k] CPD: $B = 50$ bagged exploratory trees with CPD, $d = .75$.

5.2 Individual Misclassification Rates

Prediction votes for aggregated predictors may be used to summarize the strength of individual predictions and possibly reveal errors in diagnosis. For the two-class leukemia dataset, Figure 2 displays plots of the proportions of correct classifications and three number summaries (median, lower, and upper quartiles) of the boosting prediction votes (PVs) and Golub et al. (1999) prediction strengths (PSs) for each tumor sample over the 200 LS/TS runs. The qualitative correspondence between PVs and proportions of correct classifications suggests that PVs are good indicators of a predictor's ability to correctly classify a particular tumor sample. The Golub PSs seem to be highly variable and conservative in comparison to the proportions of correct predictions, perhaps because the "voting" is over genes rather than over predictors as in PVs. Furthermore, the summands in the PSs involve expression levels for individual observations, whereas the summands in PVs involve weights, which are computed using the entire learning set. Similar results were observed for bagging PVs and other datasets (see the web supplement).

5.2.1 Lymphoma. For the lymphoma dataset, two tumor samples tended to be difficult to classify and had small prediction votes (see the web supplement for more details). The first observation (index 1, CLL-70; lymph node) is a B-CLL case, but the mRNA sample was prepared from a lymph node biopsy specimen rather than from peripheral blood cells as for other B-CLL cases. This tumor sample tended to be classified as an FL case, perhaps reflecting tissue sampling. The other observation (index 39, DLCL-0042) is believed to be a DLBCL case and tended to be classified as an FL case. The FL cases were generally harder to classify and had smaller prediction votes than tumor samples from other classes; this is likely due to the fact that the FL class only has nine observations.

5.2.2 Leukemia. For the two-class leukemia dataset, three tumor samples tended to be difficult to classify and had small prediction votes, as indicated in Figure 2. Two of these are thought to be AML cases (indices 28 and 66, corresponding to indices 48 and 72 in Figure 2) and the other is a T-cell ALL case (index 67, corresponding to index 47 in Figure 2). Samples 66 and 67 were part of the test set in the Golub et al. study and had low prediction strengths of .27 and .15. Observation 28 was part of the learning set and had a prediction strength of .44 in the Golub et al. cross-validation study.

5.2.3 NCI60. The overall performance of the predictors was much worse for the NCI 60 dataset than for the other two datasets. This is probably due to the small class sizes and the heterogeneity of some of the classes (e.g., breast cancer and NSCLC). Certain classes were easier to predict than others (e.g., colon cancer, leukemia, and melanoma), and tumor samples from these classes tended to have strongly correlated expression profiles. In addition, the triplicate leukemia (K562) and breast cancer (MCF7) samples were strongly correlated, suggesting good reproducibility of the experimental procedure (see the web supplement for more details).

5.3 Gene Selection

In general, for the lymphoma and leukemia datasets, increasing the number of genes (up to $p = 200$) did not affect greatly the performance of the various predictors. However, for the NCI60 dataset, the error rates were generally lower for $p = 200$; for instance, for DLDA, the median error rate was .33 with 200 genes and .38 with 30 genes. This is probably due to the larger number of classes and to the fact that with a

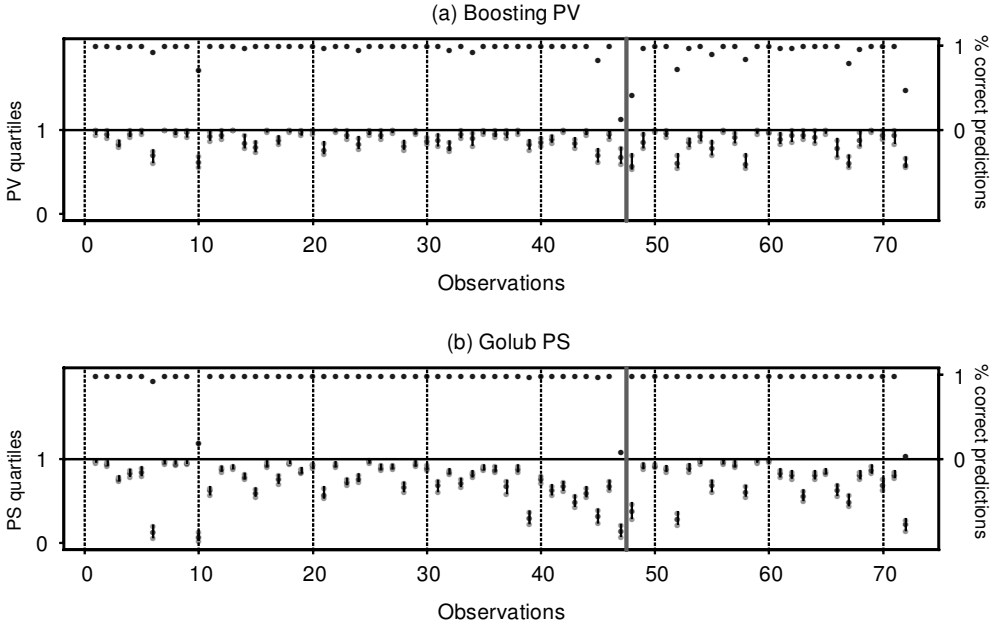

Figure 2. Leukemia Dataset, Two Classes: (a) Prediction Votes and (b) Prediction Strengths. Plots of percentages of correct predictions and three number summaries (median, lower, and upper quartiles) of prediction votes and prediction strengths for each tumor sample. (a) Displays results for CART with boosting, and (b) displays results for the weighted gene voting scheme of Golub et al. (1999). The observations are ordered by class: ALL followed by AML. Classifiers were built using $p = 40$ genes, and results are summarized over $N = 200$ LS/TS runs.

small p, a crude BW criterion is unable to identify genes that discriminate between all of the classes. Decreasing the number of genes to $p = 10$ resulted in an improved performance of FLDA. This improved performance of FLDA was even more pronounced with the "smarter" BW criterion. The performances of DLDA and DQDA were not very sensitive to the number of predictor variables, although they improved slightly with an increasing number of variables. We found aggregated predictors to be least sensitive to the number of genes used; omitting the preliminary variable selection step did not significently affect their performance (see the web supplement).

6. DISCUSSION

We have compared the performance of different discrimination methods for the classification of tumors using gene expression data from three recent studies. The main conclusion for these datasets is that simple classifiers such as DLDA and NN performed remarkably well compared with more sophisticated ones, such as aggregated classification trees. Although the lymphoma and leukemia datasets did not pose very difficult prediction problems, the NCI 60 dataset was more challenging because of the larger number of classes and the small learning set.

In the main comparison, with an intermediate number of genes selected according to a crude BW criterion, NN classifiers and DLDA had the lowest error rates, whereas FLDA had the highest. CART-based classifiers had intermediate performance, with aggregated classifiers being more accurate than a single tree. The greatest gains from aggregation were obtained by bagging with CPD and boosting. The improvement of CPD over standard bagging (i.e., nonparametric bootstrap) may be due to the fact that CPD deals with the discreteness of the sampling space by sampling from a smoothed version of the empirical cdf. For the datasets considered here, the degree of smoothing was fairly high ($d = .75$). The lack of accuracy of FLDA is likely due to the poor estimation of covariance matrices with a small learning set and a fairly large number of genes p. Indeed, decreasing the number of genes resulted in improved performance of FLDA. Also, ignoring correlations between genes as in DLDA produced lower misclassification rates than more sophisticated classifiers. For the binary class leukemia dataset, DLDA performed better than the related weighted gene voting scheme of Golub et al. (1999).

As mentioned earlier, a number of preprocessing decisions were beyond our control for these publicly available datasets. Although these decisions could in principle have a large impact on downstream analyses, we believe that the comparison of the predictors was fair, because they were applied to the same datasets. In addition, the lymphoma cDNA microarray dataset and the leukemia Affymetrix dataset were obtained by very different technologies and preprocessing steps. The similar behavior of the predictors on these two datasets leads us to believe that the results should be similar, at least qualitatively, for different preprocessing methods. Imputation of missing values is another important question that we have addressed only briefly. For CART predictors, which can deal with missing values, the prediction results were very similar for the imputed and nonimputed datasets.

Misclassification rates for the different classifiers were estimated based on random divisions of each dataset into a learning set comprising two-thirds of the data and a test set comprising one-third of the data (2:1 scheme). One needs to distinguish between two tasks: estimating misclassification rates, that is, estimating the probability that a given classifier will misclassify a new sample drawn from the same distribution as the learning set (also called generalization error), and comparing the accuracy of two or more classifiers (see Ripley 1996, chap. 2). The second task, which is our main concern here, is rather easier as classifiers are compared using the same test set. A 2:1 scheme was chosen rather than the perhaps more standard 9:1 scheme in the machine learning literature, because for our datasets the latter scheme resulted in very small test sets and more difficult discrimination between the classifiers due to the discreteness of the error rates. If our main concern was to estimate generalization error, then a 2:1 scheme would be wasteful of scarce data, which could otherwise be used for training. Also, one would need much larger datasets to get reasonably accurate estimates of generalization error.

Factors other than accuracy contribute to the merits of a given classifier. These include simplicity and insight gained into the predictive structure of the data. DLDA is easy to implement and had remarkably low error rates in our study, but it ignores correlations between predictor variables, that is, between expression levels for different genes. These correlations are biological realities, and when more data become available we may find that ignoring them is problematic. Also, LDA (with a diagonal or an arbitrary covariance matrix) cannot handle interactions between predictor variables. Gene interactions are important biologically and may contribute to class distinctions; ignoring them is not desirable. NN classifiers are simple and intuitive and had low error rates compared to more sophisticated classifiers. Although they can handle interactions between genes, they do so in a "black box" way and give very little insight into the structure of the data. In contrast, classification trees can exploit and reveal interactions between genes; they are also easy to interpret and yield information on the relationship between predictor variables and responses by performing stepwise variable selection. The main problem of single classification trees is that they tend to be unstable. Aggregation (bagging or boosting) can be used to greatly improve their accuracy. A useful byproduct of aggregated trees are PVs, which can be used to assess the confidence of predictions for individual observations. We have looked at PVs only in a qualitative manner; it would be interesting to carry out a more quantitative analysis and explore the use of thresholds for making or not making a particular prediction and for identifying errors in diagnosis. Note that the conclusions reached in our study were based on a comparison of classifiers on very small datasets by machine learning standards, that is, very small n. As more data become available, one can expect an improvement in the performance of aggregated tree classifiers, because trees should be able to correctly identify interactions. We may also be able to use these methods to gain a better understanding of the predictive structure of the data. Although some simplicity is lost by aggregating trees, aggregation may be used as part of a variable selection

approach (Fridlyand 2001). Another issue, which we have not explored and which is important in the classification of tumors, is the ability of a predictor to incorporate prior knowledge on the mRNA samples when such information is available.

Our study did not include certain popular classifiers from the field of machine learning, such as neural networks (Ripley 1996) and support vector machines (SVMs) (Vapnik 2000). We deliberately chose to look at simple predictors, which require little training. Although SVMs have been successfully applied to some problems (e.g., handwritten digit recognition), they require more training (e.g., choice of kernel function K and scale factor λ) than the predictors considered here. Also, the generalization of SVMs to more than two classes is not obvious. We are aware of a few applications of SVMs to gene expression data. SVMs were applied to the ALL/AML data, but did not improve over a simple NN or DLDA classifier (Chow, Moler, and Mian 2001). In another application, Brown et al. (2000) used SVMs to classify genes rather than mRNA samples. They considered only binary classification (i.e., each class versus its complement) and found that SVMs outperformed unaggregated classification trees and FLDA (the two worst predictors in our study). We looked into applying logistic discrimination and a perceptron classifier (Ripley 1996) to the three datasets, but our preliminary runs were not encouraging. For logistic discrimination, we encountered the well-known situation of infinite parameter estimates for perfect linear separation of the classes on the learning set. We then considered Rosenblatt's perceptron learning rule, which is specifically designed for linearly separable classes. The perceptron predictor did not generalize well and had disappointing test set error rates. We have not considered more sophisticated perceptron algorithms or penalized logistic regression, which may provide gains in accuracy. In a different type of study related to the NCI 60 drug screen, Koutsoukos et al. (1994) compared LDA, NN, and neural networks in terms of their ability to classify the 141 drug compounds into 6 groups. Our conclusions are consistent with theirs.

A very important issue that remains to be addressed is the identification of "marker" genes for tumor classes, that is, variable selection. For the purpose of comparing prediction methods, genes were selected using a simple BW criterion. A better choice for the number of genes p might be achieved by imposing a cutoff (e.g., p value) on BW or by examining plots of cumulative sums of BW versus p. For more than two classes, a criterion like BW may not always be able to identify genes that discriminate between all of the classes (cf. improvement using the "smarter" BW criterion for the lymphoma dataset). Such a criterion also tends to identify genes that are highly correlated and does not reveal interactions between genes. As sample sizes increase, one should consider methods that can exploit and discover interactions between genes (Fridlyand 2001). However, with any variable selection approach, one must be aware of the issue of statistical versus biological significance. A purely statistical approach may identify genes that reflect tissue sampling as opposed to biologically interesting and possibly unknown differences between the various tumors.

[Received June 2000. Revised July 2001.]

REFERENCES

Alizadeh, A. A., Eisen, M. B., Davis, R. E., Ma, C., Lossos, I. S., Rosenwald, A., Boldrick, J. C., Sabet, H., Tran, T., Yu, X., Powell, J. I., Yang, L., Marti, G. E., Moore, T. Jr, J. H., Lu, L., Lewis, D. B., Tibshirani, R., Sherlock, G., Chan, W. C., Greiner, T. C., Weisenburger, D. D., Armitage, J. O., Warnke, R., Levy, R., Wilson, W., Grever, M. R., Byrd, J. C., Botstein, D., Brown, P. O., and Staudt, L. M. (2000), "Different Types of Diffuse Large B-Cell Lymphoma Identified by Gene Expression Profiling," *Nature*, 403, 503–511.

Alon, U., Notterman, N. B. D. A., Gish, K., Ybarra, S., Mack, D., and Levine, A. J. (1999), "Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays," *Proceedings of the National Academy of Science*, 96, 6745–6750.

Barnard, M. (1935), "The Secular Variations of Skull Characters in Four Series of Egyptian Skulls," *Annals of Eugenics*, 6, 352–371.

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140.

——— (1998a), "Arcing Classifiers," *The Annals of Statistics*, 26, 801–824.

——— (1998b), "Using Convex Pseudodata to Increase Prediction Accuracy," Technical Report 513, Berkeley, University of California, Department of Statistics.

Breiman, L., Friedman, J. H., Olshen, R., and Stone, C. J. (1984), *Classification and Regression Trees*, Belmont, CA: Wadsworth.

Brown, M. P. S., Grundy, W. N., Lin, D., Cristianini, N., Sugnet, C. W., Furey, T. S. Jr., M. A., and Haussler, D. (2000), "Knowledge-Based Analysis of Microarray Gene Expression Data by Using Support Vector Machines," *Proceedings of the National Academy of Science*, 97, 262–267.

Chow, M. L., Moler, E. J., and Mian, I. S. (2001), "Identifying Marker Genes in Transcription Profiling Data Using a Mixture of Feature Relevance Experts," *Physiological Genomics*, 5(2), 99–111.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997), "Exploring the Metabolic and Genetic Control of Gene Expression on a Genomic Scale," *Science*, 278, 680–685.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998), "Cluster Analysis and Display of Genome-Wide Expression Patterns," *Proceedings of the National Academy of Science*, 95, 14863–14868.

Fisher, R. A. (1936), "The Use of Multiple Measurements in Taxonomic Problems," *Annals of Eugenics*, 7, 179–188.

Fix, E., and Hodges, J. (1951), "Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties," Technical report, U.S. Air Force, School of Aviation Medicine.

Freund, Y., and Schapire, R. E. (1997), "A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences*, 55, 119–139.

Fridlyand, J. (2001), "Resampling Methods for Variable Selection and Classification: Applications to Genomics," Ph.D. thesis, University of California, Berkeley, Dept. of Statistics.

Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., Coller, H., Loh, M., Downing, J. R., Caligiuri, M. A., Bloomfield, C. D., and Lander, E. S. (1999), "Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring," *Science*, 286, 531–537.

Koutsoukos, A. D., Rubinstein, L. V., Faraggi, D., Simon, R., Kalyandrug, S., Weinstein, J. N., Kohn, K., and Paull, K. (1994), "Discrimination Techniques Applied to the NCI In Vitro Anti-Tumour Drug Screen: Predicting Biochemical Mechanism of Action," *Statistics in Medicine*, 13, 719–730.

Lockhart, D. J., Dong, H. L., Byrne, M. C., Follettie, M. T., Gallo, M. V., Chee, M. S., Mittmann, M., Wang, C., Kobayashi, M., and Horton, H. (1996), "Expression Monitoring by Hybridization to High-Density Oligonucleotide Arrays," *Nature Biotechnology*, 14, 1675–1680.

Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979), *Multivariate Analysis*, San Diego: Academic Press.

McLachlan, G. J. (1992), *Discriminant Analysis and Statistical Pattern Recognition*, New York: Wiley.

Perou, C. M., Jeffrey, S. S., van de Rijn, M., Rees, C. A., Eisen, M. B., Ross, D. T., Pergamenschikov, A., Williams, C. F., Zhu, S. X., Lee, J. C. F., Lashkari, D., Shalon, D., Brown, P. O., and Botstein, D. (1999), "Distinctive Gene Expression Patterns in Human Mammary Epithelial Cells and Breast Cancers," *Proceedings of the National Academy of Science*, 96, 9212–9217.

Pollack, J. R., Perou, C. M., Alizadeh, A. A., Eisen, M. B., Pergamenschikov, A., Williams, C. F., Jeffrey, S. S., Botstein, D., and Brown, P. O. (1999), "Genome-Wide Analysis of DNA Copy-Number Changes Using cDNA Microarrays," *Nature Genetics*, 23, 41–46.

Ripley, B. D. (1996), *Pattern Recognition and Neural Networks*, Cambridge, U.K.: Cambridge University Press.

Ross, D. T., Scherf, U., Eisen, M. B., Perou, C. M., Spellman, P., Iyer, V., Jeffrey, S. S., de Rijn, M. V., Waltham, M., Pergamenschikov, A.,

Lee, J. C. F., Lashkari, D., Shalon, D., Myers, T. G., Weinstein, J. N., Botstein, D., and Brown, P. O. (2000), "Systematic Variation in Gene Expression Patterns in Human Cancer Cell Lines," *Nature Genetics*, 24, 227–234.

The Chipping Forecast (January 1999), Vol. 21, Supplement to *Nature Genetics.*

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D., and Altman, R. B. (2001), "Missing Value Estimation Methods for DNA Microarrays," *Bioinformatics*, 17, 520–525.

Vapnik, V. N. (2000), *The Nature of Statistical Learning Theory* (2nd ed.), New York: Springer.

Yang, Y. H., Buckley, M. J., Dudoit, S., and Speed, T. P. (2002), "Comparison of Methods for Image Analysis on c(DNA) Microarray Data," *Journal of Computational and Graphical Statistics*, 11.

Yang, Y. H., Dudoit, S., Luu, P., and Speed, T. P. (2001), "Normalization for cDNA Microarray Data," in *Microarrays: Optical Technologies and Informatics*, eds. M. L. Bittner, Y. Chen, A. N. Dorsel, and E. R. Dougherty.