Using Nuclear Morphometry to Discriminate the Tumorigenic Potential of Cells: A Comparison of Statistical Methods

Pamela Wolfe,[1] James Murphy,[1] John McGinley,[2] Zongjian Zhu,[1] Weiqin Jiang,[1] E. Brigitte Gottschall,[2] and Henry J. Thompson[1]

[1]Cancer Prevention Laboratory, Colorado State University, Fort Collins, Colorado and [2]Departments of Biometrics and Occupational Medicine, National Jewish Medical and Research Center, Denver, Colorado

Abstract

Despite interest in the use of nuclear morphometry for cancer diagnosis and prognosis as well as to monitor changes in cancer risk, no generally accepted statistical method has emerged for the analysis of these data. To evaluate different statistical approaches, Feulgen-stained nuclei from a human lung epithelial cell line, BEAS-2B, and a human lung adenocarcinoma (non-small cell) cancer cell line, NCI-H522, were subjected to morphometric analysis using a CAS-200 imaging system. The morphometric characteristics of these two cell lines differed significantly. Therefore, we proceeded to address the question of which statistical approach was most effective in classifying individual cells into the cell lines from which they were derived. The statistical techniques evaluated ranged from simple, traditional, parametric approaches to newer machine learning techniques. The multivariate techniques were compared based on a systematic cross-validation approach using 10 fixed partitions of the data to compute the misclassification rate for each method. For comparisons across cell lines at the level of each morphometric feature, we found little to distinguish nonparametric from parametric approaches. Among the linear models applied, logistic regression had the highest percentage of correct classifications; among the nonlinear and nonparametric methods applied, the Classification and Regression Trees model provided the highest percentage of correct classifications. Classification and Regression Trees has appealing characteristics: there are no assumptions about the distribution of the variables to be used, there is no need to specify which interactions to test, and there is no difficulty in handling complex, high-dimensional data sets containing mixed data types. (Cancer Epidemiol Biomarkers Prev 2004;13(6):976–88)

Introduction

The size, shape, and chromatin pattern of nuclei, which is operationally defined as nuclear morphometry of malignant cells, are known to differ from that observed in nonmalignant cell nuclei. Based on these differences, efforts have been made to use the quantitative assessment of nuclear morphometry to enhance diagnostic and prognostic efforts of pathologists (1-6). It also has been observed that more subtle differences in nuclear morphometry are observed in histologically normal-appearing cells in areas that are peripheral to a malignant lesion (7, 8). These observations have led to the hypothesis that such changes occur prior to the emergence of clinically detectable disease and that nuclear morphometry can be used as a surrogate endpoint biomarker for estimating an individual's risk for cancer (9-12).

While adaptation of quantitative morphometric analysis as a clinical tool has been discussed in the literature for over 30 years, the usefulness of this approach has not received widespread acceptance. One reason for this is that there has been limited evidence that this approach can provide stable, specific, and selective algorithms for use in diagnosis, prognosis, or risk assessment. This is reflected in the literature by less than optimal classification rates and the lack of consistency among reports in the morphometric variables selected for use in diagnostic, prognostic, or disease risk algorithms (10, 11, 13, 14). The situation has been further complicated by the emergence in the last few years of programs that compute an increasing number of morphometric parameters derived from the same basic set of image data, which raises the question of whether the additional parameters are providing new information or simply additional measurement variables. If it is the latter, data reduction should be among the steps taken prior to modeling.

Various statistical methods have been applied to the analysis of morphometric data, but our search did not reveal any work in comparative discrimination methods. The number of morphometric variables is not so substantial as to discourage attempts to evaluate the data one parameter at a time and not yet enough to suggest that the application of artificial intelligence, or machine learning techniques, would be the method of choice. The data are not multinormal, but the performance of classic statistical tests, such as Student's *t*, is optimal when the assumption of normality holds. The power of nonparametric methods depends on just how far from normal the distribution strays; nonparametric multivariate methods require of the analyst some statistical savvy and patience with trial and error if they are to be applied appropriately. Artificial intelligence

Received 10/6/03; revised 12/23/03; accepted 2/12/04.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked advertisement in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Pamela Wolfe, Cancer Prevention Laboratory, Colorado State University, 111 Shepardson Building, 1173 Campus Delivery, Fort Collins, CO 80523. Phone: 303-730-3236; Fax: 970-491-3542. E-mail: wolfep@earthlink.net

algorithms tend to work best with high-dimensional data sets, and neural networks, in particular, operate in a black box, producing an answer but failing to show their work. In an effort to evaluate the relative merits of a variety of statistical methods using an experimental approach that was simple and designed to limit individual variation among cells being evaluated, this study was performed using cells grown in culture, the biological characteristics of which relative to malignancy were known.

R.A. Fisher's seminal article in the *Annals of Eugenics* (1936) described a linear discriminant function for the taxonomy of three *Iris* sp. based on four parameters: sepal length and width, petal length and width. Although the method he described was general and did not assume normality, it has since been shown to perform optimally when that assumption holds. Comparison of linear discriminant analysis (LDA) with logistic regression (LR) has shown that when the data are not multivariate normal, LR is a more reliable classifier, and more so as the distributions diverge from normal. Some examples are Efron's (15) theoretical treatment, an application to more than two groups by Bull and Donner (16), and a simulation study that compared the performance of multiple group linear discriminant function, rank-based linear discriminant function, kernel density (KD) function, and multiple logistic function by Barón (17). The advent of high-throughput devices that is revolutionizing biomedical research is also challenging statisticians; an empirical comparison of the performance of classic discriminant analysis with several computer-intensive, nonparametric techniques applied to genomic data was published in the *Journal of the American Statistical Society* in 2002 (18). A theoretical treatment of discrimination and classification is given by Hand (19); computer-intensive machine learning techniques are described by Hastie et al. (20).

This article is organized into a Materials and Methods section, which describes the lung cell culture and image acquisition; a description of the Statistical Methods we applied to the image data; the Results of the classification analysis; and a Discussion of the results with an emphasis on the strengths of those we deem most appropriate for classification based on morphometric data.

Materials and Methods

Cell Culture. Two human bronchial epithelial cell lines were selected for this work: BEAS-2B, a cell line derived from normal bronchial epithelial cells immortalized with adenovirus, and NCI-H522, a human lung adenocarcinoma (non-small cell) cancer cell line. The characteristics of these cell lines are summarized in Table 1. Cells were obtained from American Type Culture Collection (Manassas, VA). Cells were grown at 37°C in a humidified incubator with 5% CO_2. BEAS-2B cells were cultured in LHC-9 serum-free medium (Clonetics, Walkersville, MD) with supplements (Clonetics) of 0.5 ng/mL recombinant epidermal growth factor, 500 ng/mL hydrocortisone, 0.005 mg/mL insulin, 0.035 mg/mL bovine pituitary extract, 500 nmol/L ethanolamine, 500 nmol/L phosphoethanolamine, 0.01 mg/mL transferrin, 6.5 ng/mL 3,3′,5-triiodothyronine, 500 ng/mL epinephrine, 0.1 ng/mL retinoic acid, and trace elements. NCI-H522 cells were cultured in RPMI 1640 (American Type Culture Collection) with 2 mmol/L L-glutamine (Life Technologies, Inc., Rockville, MD) adjusted to contain 1.5 g/L sodium bicarbonate (Sigma Chemical Co., St. Louis, Missouri), 4.5 g/L glucose, 10 mmol/L HEPES (Sigma Chemical), and 1.0 mmol/L sodium pyruvate (Sigma Chemical) plus 10% fetal bovine serum (American Type Culture Collection).

Harvesting Cells. After 48 hours in culture, cells reached 80% to 90% confluence. At this stage, cells were rinsed and the culture dish bottom was scraped in PBS. The resulting suspension of cells was spun at 1000 rpm for 5 minutes and fixed in 5 mL Saccomanno fluid. Cells were subsequently cytospun onto glass slides, postfixed in 10% neutral-buffered formalin for 30 minutes, rinsed in deionized water for 5 minutes, and allowed to air dry prior to Feulgen staining (21).

Image Acquisition and Analysis. The nuclear morphometry of the cells was evaluated by computer-assisted image analysis using a CAS-200 image analysis system (Becton-Dickinson, San Jose, CA) equipped with CellSheet version 2.0 software (Bacus Laboratories, Lombard, IL). For analysis, 100 diploid cells from each cell line were selected; no S-phase, tetraploid, or aneuploid cells were evaluated. In addition, irregularly

Table 1. Summary of the characteristics of two lung epithelial cell lines

Criteria	NCI-H522	BEAS-2B
Origin	Derived from a human primary non-small cell lung adenocarcinoma.	Derived from epithelial cells isolated from normal human bronchial epithelium obtained from autopsy of an individual without grossly apparent lung cancer.
Tumorigenic	Yes. Mutations in p53 mutation at codon 191 and K-*ras* at codon 12.	No. The cells were immortalized by infection with an adenovirus hybrid (Ad12SV40) and cloned.
Morphology	Epithelial. Adenocarcinoma and hypotriploid human cell line with the modal chromosome number of 53 in 68% of cells counted. The polyploid cells occurred at 3.0%.	Epithelial. Cells retain the ability to undergo squamous differentiation.
Growth properties	Adherent	Adherent
Usage	Research only.	Used to screen chemical and biological agents that induce or affect differentiation and/or carcinogenesis.

shaped diploid nuclei were excluded from analysis. Cells were selected for acquisition via digital filter. Cell images were screened by a technician following acquisition for quality control purposes. Overlapping or oddly shaped nuclei were deleted prior to nuclear morphometric analysis. Parameters derived from the size, shape, and absorbance of each nucleus were measured. In addition to these 12 measurements, 6 general texture and 21 Markovian texture features of the nucleus were evaluated (Table 2). Markovian texture features characterize variations between adjacent pixels in an image based on mathematical algorithms using a gray-level transition probability matrix. Markovian texture data provide a quantitative assessment of fine changes in chromatin structure classically used by the pathologist for grading and staging of lesions. However, the use of computer-assisted image analysis removes the subjectivity and greatly improves the sensitivity to detect subtle early changes in chromatin structure that are associated with cancer. All 39 features are listed in Table 2 and are described in detail at http://cpl.colostate.edu/morph. Differences measured by morphometric analysis are generally subtle and are frequently not detected on visual inspection of stained cells. Images of the cells measured in this study and of the cells that were typically misclassified statistically are provided at http://cpl.colostate.edu/morph.

Statistical Methods

Descriptive Statistics. Graphic description of the data consisted of density plots for each morphometry parameter and pairwise scatter plots for a subset of parameters that were selected by at least one stepwise procedure. Mean and SD for each parameter by cell line also were computed. A customized z-score for each parameter in the cancer cell line was computed based on the mean and SD of each parameter in the normal cell line (22). The z-score for the ith feature in the jth cell line for the kth observation can be written as $z_{ijk} = (x_{ijk} - \bar{x}_i) / \sigma_l$, where 1 designates the reference group, the nontumorigenic BEAS-2B cell line. This customized z-score measures the direction and distance from the mean of the reference group in SD units. The mean z-scores for the reference group are all 0 by definition.

Classification. The following notation and terms are used throughout the discussion of classification methods. Let $\boldsymbol{X}$ be a P-dimensional measurement space containing all possible ordered vectors of morphometry features; for our data, the first coordinate, x_1, is nuclear area; the second, x_2, is PgDNA; . . .; the last, x_p, is triangular symmetry (refer to Table 2). A classifier d($\mathbf{x}$) is a function defined on $\boldsymbol{X}$ that will, for J cell types, partition the measurement space into J pairwise disjoint subsets $A_1, A_2, \ldots, A_J$; d($\mathbf{x}$) assigns each vector $\mathbf{x} = (x_1, x_2, \ldots, x_p)$ from $\boldsymbol{X}$ to exactly one subset, A_j. For every case with a measurement vector $\mathbf{x}$ in A_j, the class assignment is j. Classifiers are constructed from a learning sample consisting of N vectors, where $\mathbf{x_n} = (x_{n1}, x_{n2}, \ldots, x_{np})$ is drawn from $\boldsymbol{X}$ with known class membership, y_n. Define a set of N cases with an indicator, y, for class membership to be a learning set $\boldsymbol{L} = \{(x_1, y_1), (x_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$. The error rate computed by counting the misclassified elements in $\boldsymbol{L}$ for a classifier developed on $\boldsymbol{L}$ is the resubstitution rate and is overly optimistic. The validity, or performance, of a classifier is estimated from a test set, $\boldsymbol{T} = (x_1, x_2, \ldots, x_M)$, where y_m is known but used only after the classifier has been applied to compute the misclassification rate. In our data set, $N = 200$, $P = 39$, and $J = 2$; Y is coded 1 for the cancer cell line, 0 otherwise.

LDA, quadratic discriminant analysis, and LR belong to a class of methods that estimate a discriminant function, d($\mathbf{x}$), and classify an observation $\mathbf{x}$ based either on threshold values of d($\mathbf{x}$) or on the minimum over j of $\mathrm{d}_j(\mathbf{x})$, which is the distance between $\mathbf{x}$ and the center of group j.

Table 2. Feature categories of nuclear morphometric parameters

Size	Shape	General Texture	Markovian Texture
Nuclear area	Perimeter	Run length	Angular second moment
DNA content (pg)	Elongation	Configurable run length	Contrast
Absorbance	Min diameter	Valley, slope, peak	Correlation
Average absorbance	Max diameter	SD	Difference moment
Sum absorbance	Cell Feret X		Inverse difference moment
	Cell Feret Y		Sum average
	Shape		Sum variance
			Sum entropy
			Entropy
			Difference variance
			Difference entropy
			Information measure A
			Information measure B
			Maximal correlation coefficient
			Coefficient of variation
			Peak transitional probability
			Diagonal variance
			Diagonal moment
			Second diagonal moment
			Product moment
			Triangular symmetry

Suppose the class probability densities are multivariate normal, that is, $\mathbf{x}|y_j \sim N_p(\mu_j, \Sigma_j)$ and π_j is the proportion of observations in class j. Then, a discriminant function that achieves maximum separation of means is given by $\mathrm{d}_j(\mathbf{x}) = -(1/2)\log\Sigma_j - (1/2)(\mathbf{x} - \mu_j)^T\Sigma_j^{-1}(\mathbf{x} - \mu_j) + \log \pi_j$ and its application is called quadratic discriminant analysis. The second term, $(x - \mu_j)^T\Sigma_j^{-1}(x - \mu_j)$, is the Mahalanobis distance from $\mathbf{x}$ to the center of the class. If $\Sigma_j = \Sigma$ for all j, the discriminant function simplifies to $\mathrm{d}_j(\mathbf{x}) = \mathbf{x}^T\Sigma^{-1}\mu_j - (1/2)\mu_j^{\mathrm{T}}\Sigma_j^{-1}\mu_j + \log \pi_j$, which is linear in $\mathbf{x}$ and is called LDA. The population parameters are rarely known and sample estimates for μ and Σ are substituted into the equations in the development and application of $\mathrm{d}_j(\mathbf{x})$. Although LDA and quadratic discriminant analysis perform optimally with multinormal random variables, in practice, the normality assumption is often ignored. For a detailed discussion of discriminant analysis under the assumption of normality, see Morrison (23).

LR (24) models the posterior probability that a case is in one of two or more classes conditional on $\mathbf{x}$. Let $\pi(\mathbf{x}) = E(Y|\mathbf{x})$, the expected value of Y conditional on $\mathbf{x}$. The LR model is specified: $\pi(\mathbf{x}) = \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)/[1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p)]$ which is nonlinear in $\mathbf{x}$.

The logit transformation $g(\mathbf{x}) = \ln[\pi(\mathbf{x}) / (1 - \pi(\mathbf{x}))] = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$ gives a function that is linear in $\mathbf{x}$ and ranges from $-\infty$ to $+\infty$. Estimates for $\pi(\mathbf{x})$ can now be seen as an example of $\mathrm{d}(\mathbf{x})$, described above, and when $J = 2$ classes, we assign $\mathbf{x}$ to class 2 if $\pi(\mathbf{x}) > c$ and class 1 if $\pi(\mathbf{x}) \leq c$. A common choice for c is 0.5, but this may not maximize sensitivity and specificity. While this model is still linear in $\mathbf{x}$, the distributional assumption is on Y (binomial) rather than $\mathbf{x}$, as it was in LDA and quadratic discriminant analysis; the model is easily extended to multinomial outcomes.

Binary tree structured classifiers begin by splitting the sample space into two partitions; each partition, or node, is split into two more partitions, and so on until further splitting cannot improve the performance of the classifier. The construction requires three elements: a rule for determining the splits, a rule for stopping splitting, and a rule for class assignment to terminal nodes or leaves. The criterion for splitting is a reduction in impurity, a function of the class representation at a node, which is at its maximum value when all classes are equally represented at a node and at its minimum when a node contains only one class. The size of the tree can be limited in two ways, stopping rules and pruning. Stopping rules are based on a combination of impurity and the number of cases at a node. Pruning can be based on a cost complexity algorithm or the misclassification rate. Class assignment is determined by plurality at a node. The method we used was Classification and Regression Trees (CART), described by Breiman et al. (25) and implemented in the tree() function in S-Plus.

K nearest neighbor (KNN) clustering is a memory-based classifier and imposes no distributional assumptions on the data. The method requires only a metric $\|\mathbf{xi}\|$ on $\mathbf{X}$ and the specification of K, the number of neighbors that contribute to the decision on class assignment of $\mathbf{x}$. A range of values for K is tested by cross-validation. Using the K that minimizes the misclassification rate, the procedure is to select the KNN of $\mathbf{x_i}$ based on the metric and classify $\mathbf{x_i}$ by majority vote. The resulting partition of $\mathbf{X}$ is not linear. The method is sensitive to the choice of metric and provides no information about the structure of the data. We selected the Mahalanobis distance as the metric, after some trial and error, and set $K = 5$. KNN is implemented in SAS as an option in PROC DISCRIM and in the R function knn().

The KD method, like KNN, generates a nonlinear partition of the data, is nonparametric, and provides little information about the structure of the data. The method incorporates a specific weighting function, $K_j(\|\mathbf{x}\|)$, the kernel, and a fixed radius, r, to estimate the probability that $\mathbf{x_i}$ belongs to group j. The posterior probability that $\mathbf{x_i}$ belongs to j is given by Bayes' rule: $p(j|\mathbf{x}) = \pi_j f_j(\mathbf{x}) / f(\mathbf{x})$, where the probability density estimate $f_j(\mathbf{x}) = n_j^{-1}\Sigma_y K_j(\mathbf{x} - \mathbf{y})$, summing over every observation $\mathbf{y}$ in group j, and $f(\mathbf{x})$ is the unconditional density, $\Sigma_u \pi_u f_u(\mathbf{x})$, summing over all groups. $\mathbf{x_i}$ is assigned to the group where $p(j|\mathbf{x})$ is greatest. The uniform kernel method is easily visualized: $p(j|\mathbf{x})$ is a function of the number of observations, y, from group j falling in a p-dimensional ellipsoid of radius r centered on $\mathbf{x_i}$. This is different from KNN only in that the radius is fixed. Other KD functions weight the distance between $\mathbf{x_i}$ and each $\mathbf{y}$ in j. Optimal classification for the lung cell culture data was given by $r = 2$ and the normal kernel: $K_j(\|\mathbf{x}\|) = [1 / c(j)]\exp(-0.5\|\mathbf{x}\|^T\Sigma_j^{-1}\|\mathbf{x}\| / r^2)$, where $c(j) = (2\pi)^{p/2}r^p|\Sigma_j^{-1}|^{1/2}$.

Other Methods. Newer methods for exploring higher-dimensional data, such as neural networks (14, 26, 27) and the learning vector quantization neural network (13), have been applied to morphometry data. We tried several specifications of a neural network and a support vector machine on this data set; the results (available from http://cpl.colostate.edu/morph) were poor. For a comprehensive discussion of statistical learning, see Hastie et al. (20).

Software. We used SYSTAT version 10 for graphics; SAS version 8.2, S-Plus 4.0, R,[3] and iMiner 1.01 were used for all other analyses. In most cases, default parameters were used: stepwise LR in SAS adds or removes variables at the 0.05 significance level and stepwise discriminant analysis adds or removes variables when their partial correlation coefficient is 0.01. The CART procedure in S-Plus uses deviance (stop splitting when impurity is <1%) and the number of observations at a node to stop splitting. We reduced the minimum number of cells per leaf for a split to 2 (the default is 10) and specified the minimum deviance (the specific measure of impurity) to be 0.001. We did a search from 1 to 10 for the number of neighbors that minimized errors for the KNN application; misclassification within the learning set was minimized with sets of five neighbors; the default metric is the Mahalanobis distance computed using the full covariance matrix. We also tested the Euclidean distance and the Mahalanobis distance computed using only the diagonal of the covariance matrix in an effort to reduce the misclassification rate. Mahalanobis distance using the full covariance matrix was the best metric. KD also required an iterative approach to the best specification.

[3] R is an open source software environment for statistical computing and graphics, similar to the S system, which was developed at Bell Laboratories by J. Chambers et al. A variety of sophisticated analysis packages can be downloaded from http://www.r-project.org. Analysis packages for S-Plus are available at http://lib.stat.cmu.edu.

Criteria for Comparison of Statistical Methods. The purpose of classification is prediction and understanding; while these goals are not mutually exclusive, not every method reveals the structure of the data. A method that elucidates structure or patterns in a data set, other things equal, would be preferred over one that does not. One straightforward criterion for selecting one method over another is prediction accuracy or the proportion of cases that are misclassified. Because misclassification rates do not indicate the direction of the errors (the false-positive and false-negative rates), sensitivity and specificity are also informative; in a particular application, it may be more important to minimize error in one direction than the other.

The validity of a classifier should be assessed by computing the misclassification rate on an independent sample of data. For large data sets, this is accomplished by holding out half or a third of the data to be used as a validation set, $\boldsymbol{T}$. For smaller data sets such as ours, where a more parsimonious use of data is required, v-fold cross-validation works well (25). Briefly, the cases in $\boldsymbol{L}$ are divided randomly into V subsets of equal size designated $\boldsymbol{L}_1, \boldsymbol{L}_2, \ldots, \boldsymbol{L}_V$. The classifier is constructed on $\boldsymbol{L} - \boldsymbol{L}_v$ and tested on $\boldsymbol{L}_v$ for all v. The resulting misclassification rates are averaged over V, and the final classifier is constructed on $\boldsymbol{L}$. Cross-validation is parsimonious with data; each case is used to construct the classifier and each case is used once in a test sample. For analyses in SAS and R, the data were partitioned into 10 sets of 20 observations, balanced between cell lines. The same partition was used for each classification method, and for stepwise methods, variable selection was done on each partition to avoid the bias that would be introduced by using any information from the test set, $\boldsymbol{L}_{v.}$

Misclassification has two components: sensitivity and specificity. In the context of our data, sensitivity is the ability of a classifier to correctly identify cells in the cancer cell line and specificity is the ability of a classifier to correctly identify cells in the normal cell line. These were computed along with the misclassification rate in the cross-validation process and are reported in Table 3. Although there is no cost associated with false positives in our experiment, in many settings, this would be an additional consideration.

The misclassification rate, sensitivity, and specificity are associated with a single cut point; $\Pr(Y = 1) = 0.5$ is the default threshold for determining class membership. The receiver operating characteristic (ROC) curve shows the tradeoff between specificity and sensitivity as one or the other is increased by moving the cut point away from 0.5 in either direction. It may be possible in some applications to improve sensitivity without increasing the misclassification rate or to reduce the misclassification rate without sacrificing sensitivity. The area under the ROC curve provides an alternative description of classification accuracy that is independent of the cut point; it is calculated by comparing the estimated probabilities for membership in the event class for all possible pairs of cases in the two classes. For n_0 cases with $Y = 0$ and n_1 cases with $Y = 1$, there are n_0n_1 pairs. Count the number of times cases for which $Y = 1$ have a higher probability than cases in the $Y = 0$ class, assign 0.5 to ties, sum, and divide by n_0n_1. The area under the ROC curve was computed from the estimated probabilities for each test set, $L_{v.}$

Results

Basic descriptive statistics inform all subsequent analyses. The distribution of a representative subset of the morphometric parameters is shown in Fig. 1 (all 39 are on our Web site http://cpl.colostate.edu/morph). Many are skewed, truncated normal, or log normal, suggesting the use of statistical techniques that depend heavily on the assumption of normality should be avoided. The parameters are also highly correlated because some are

Table 3. Comparison of five classification methods for predicting cell line

Method	Variables Selected from Full Data Set	Misclassification Rate* (SD; %)	Sensitivity (%)	Specificity (%)	Area under ROC (SD)
KNN (SAS)	All with bootstrap $P < 0.20$ ($k = 5$, metric = Mahalanobis distance)	16.0 (7.4)	81	87	0.909 (0.053)
KD (SAS)	All with bootstrap $P < 0.20$ ($r = 2$, kernel = normal, metric = Mahalanobis distance)	13 (5.4)	92	83	0.952 (0.039)
Stepwise LDA (SAS)	Coefficient of variation, PgDNA, sum variance, cell Feret Y, slope, valley	11.0 (3.9)	86	92	0.959 (0.027)
Stepwise LR (SAS)	Valley, coefficient of variation, diagonal moment, sum absorbance	10.5 (6.9)	88	91	0.953 (0.032)
CART (S-Plus)	Product moment, difference variance, sum variance, contrast, peak, PgDNA	11.0 (7.4)	89	89	0.893 (0.067)

NOTE: $n = 100$ cells from BEAS-2B (normal bronchial epithelial cells) and 100 cells from NCI-H522 (non-small cell lung adenocarcinoma). Sensitivity is the correctly classified percentage of cancer cells; specificity is the correctly classified percentage of normal cells.

*Method: 10-fold cross-validation: the data set was partitioned into 10 sets of 20 observations, 10 cancer and 10 normal. Each set of 20 was used as the test set for parameters estimated using the remaining 180 data points as the learning set. The misclassification rate for the test set and its SD were calculated from the misclassification rates across the 10 models. The area under the ROC curve and its SD were calculated from the area under the ROC curve across the 10 models.

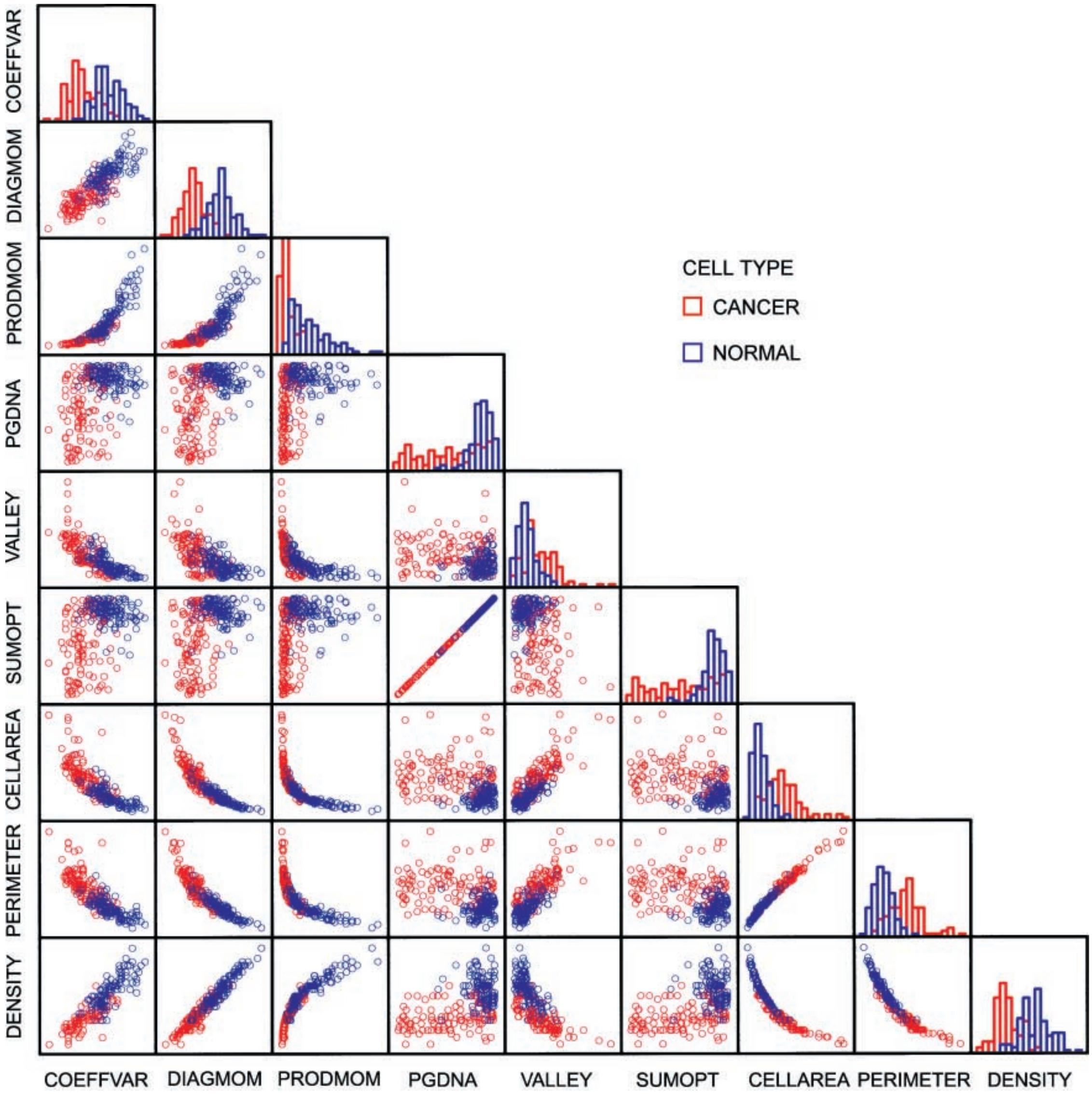

Figure 1. Pairwise correlations (*off-diagonal*) and histograms (*diagonal*) for nine features, either selected at least once by a stepwise classification method (see Table 3) or showing strong nonlinear correlation by cell line.

derived from others and have common elements. The correlation between absorbance and sum absorbance, for example, is 0.99 in this data set, while the correlation between density and perimeter is −0.98. Of the 741 possible pairwise correlations between 39 parameters, 152 (21%) have absolute value >0.90 and another 78 (11%) have absolute value between 0.80 and 0.90. Pairwise scatter plots for the subset of nine parameters are also shown in Fig. 1. Redundancy is one of the salient characteristics of these data; removing a subset of parameters from the analysis could result in little or no loss of information. We used stepwise selection with several of the discriminant procedures.

The t test approaches the problem of $J = 2$ populations by testing the null hypothesis that two independent samples were drawn from the same population. While the usefulness of a univariate t test as a classifier is limited in the multivariable setting, the basis for the formulation of the linear discriminant function is Hotelling's multivariate T^2 (19). Examining each parameter in this way does contribute to the description of the data structure and can be used for variable selection by identifying a subset of parameters that do not differ across classes. Many of the parameters in the lung cell culture data are dissimilar across the two cell lines. Among the unadjusted P values shown in Table 4, only five are >0.05. Although the t test is fairly robust to violations of the assumption of normality, the additional problem of multiple tests makes interpretation of the raw P values difficult, because the "$P < 0.05$" outcomes occur more frequently even when there are no real differences. In contrast to the familiar Bonferroni-style adjustments for multiple comparisons, resampling methods incorporate the correlation structure and distributional characteristics of the data. The step-down tests on bootstrap estimates (28, 29) of the differences in cell line means, in which it is not necessary to assume a particular distribution for the errors, provide a more reasonable estimate of the differences than the unadjusted *P* values: seven parameters are not significantly different between the cell lines. The *P* values in the column labeled ''StepBoot'' in Table 4 are adjusted for multiple comparisons. See http://cpl.colostate.edu/morph for details.

The *z*-scores for the cancer cell line data give another perspective on the extent of difference in nuclear morphometric parameters between cell lines. The *z*-scores measure the difference in means in normal cell line SD units: 26 (67%) of the scores have absolute value <1 and 9(23%) have absolute value >2. Mean *z*-scores for the cancer NCI-H522 cell line are shown in Table 4. The *z*-scores have been used in conjunction with coefficients from discriminant analysis, described below, to construct a weighed measure of morphometric differences. Bacus et al. (22) called this nuclear grade (NG).

The *z*-scores are consistently greater in absolute value for the parameters in the categories of size, shape, and general texture, all of which are more familiar to the histologist than the Markovian texture parameters. It is not surprising that the Markovian texture parameters, which are mathematical constructions, appear to carry somewhat less information. However, the feature selected first in the stepwise procedures tends to belong to the Markovian texture subset (coefficient of variation). Good separation and a high *z*-score are not synonymous. Figure 2 illustrates this: for sum absorbance (sum optical density), $z = -2.93$; for the coefficient of variation, $z = -1.87$; however, the sum absorbance values for the cancer cells lie entirely within the distribution of values for the normal cells.

Stepwise LDA selected six parameters for the classification of cells into their respective cell line categories: coefficient of variation, PgDNA, sum variance, cell Feret Y, slope, and valley, and these were selected regardless of whether the input was raw measures or *z*-scores for

Table 4. Descriptive statistics by parameter and cell line

		Cell Line		*P*		
Morphometric Category	Morphometric Feature	Normal	Cancer	Raw	StepBoot	*z*-Score
Size	Area	37.47 ± 8.02	59.80 ± 17.81	<0.0001	<0.0001	2.79
	PgDNA	7.75 ± 0.17	7.25 ± 0.46	<0.0001	<0.0001	−2.94
	Absorbance	0.21 ± 0.04	0.13 ± 0.04	<0.0001	<0.0001	−1.97
	Average absorbance	0.45 ± 0.09	0.27 ± 0.08	<0.0001	<0.0001	−1.99
	Sum absorbance	89.47 ± 1.96	83.72 ± 5.30	<0.0001	<0.0001	−2.93
Shape	Shape	13.24 ± 0.53	13.75 ± 0.64	<0.0001	<0.0001	0.96
	Perimeter	23.00 ± 2.67	29.50 ± 4.75	<0.0001	<0.0001	2.44
	Elongation	1.27 ± 0.13	1.28 ± 0.18	0.6536	0.7121	0.08
	Min diameter	6.41 ± 0.71	8.01 ± 1.31	<0.0001	<0.0001	2.25
	Max diameter	8.13 ± 0.95	10.16 ± 1.49	<0.0001	<0.0001	2.15
	Cell Feret X	7.89 ± 0.92	9.66 ± 1.46	<0.0001	<0.0001	1.93
	Cell Feret Y	7.43 ± 0.90	9.39 ± 1.59	<0.0001	<0.0001	2.18
General texture	Run length	193.20 ± 37.25	274.40 ± 54.92	<0.0001	<0.0001	2.18
	Configurable run length	33.58 ± 21.73	65.65 ± 41.45	<0.0001	<0.0001	1.48
	Valley	28.61 ± 18.45	55.07 ± 34.67	<0.0001	<0.0001	1.43
	Peak	100.69 ± 24.32	135.01 ± 44.15	<0.0001	<0.0001	1.41
	Slope	478.22 ± 98.95	689.49 ± 128.66	<0.0001	<0.0001	2.14
	SD	0.25 ± 0.08	0.11 ± 0.05	<0.0001	<0.0001	−1.69
Markovian texture	Angular second moment	0.03 ± 0.00	0.04 ± 0.01	<0.0001	<0.0001	0.76
	Correlation	0.84 ± 0.02	0.84 ± 0.04	0.4574	0.7121	0.15
	Coefficient of variation	0.48 ± 0.07	0.36 ± 0.07	<0.0001	<0.0001	−1.87
	Difference entropy	0.50 ± 0.02	0.49 ± 0.04	0.0586	0.2023	−0.37
	Entropy	1.49 ± 0.02	1.48 ± 0.04	0.0033	0.0161	−0.59
	Information measure A	−0.32 ± 0.03	−0.33 ± 0.04	0.0488	0.1730	−0.39
	Information measure B	0.66 ± 0.02	0.67 ± 0.03	0.1595	0.3383	0.29
	Maximal correlation coefficient	0.75 ± 0.05	0.77 ± 0.06	0.0389	0.1555	0.35
	Product moment	552.99 ± 356.41	125.68 ± 110.92	<0.0001	<0.0001	−1.20
	Sum variance	2411.79 ± 1556.39	545.09 ± 478.26	<0.0001	<0.0001	−1.20
	Sum average	97.94 ± 19.98	58.87 ± 16.84	<0.0001	<0.0001	−1.96
	Contrast	199.85 ± 136.78	42.35 ± 36.99	<0.0001	<0.0001	−1.15
	Diagonal moment	11.84 ± 2.01	7.85 ± 1.72	<0.0001	<0.0001	−1.98
	Diagonal variance	0.00 ± 0.00	0.00 ± 0.00	<0.0001	0.0002	0.74
	Difference moment	9.73 ± 3.38	4.20 ± 1.90	<0.0001	<0.0001	−1.64
	Difference variance	93.85 ± 62.52	21.11 ± 17.98	<0.0001	<0.0001	−1.16
	Inverse difference moment	0.36 ± 0.04	0.43 ± 0.06	<0.0001	<0.0001	1.86
	Peak transition probability	0.09 ± 0.01	0.10 ± 0.02	<0.0001	<0.0001	0.82
	Triangular symmetry	0.27 ± 0.09	0.21 ± 0.08	<0.0001	<0.0001	−0.68
	Second diagonal moment	4.86 ± 1.69	2.10 ± 0.95	<0.0001	<0.0001	−1.64
	Sum entropy	1.15 ± 0.01	1.15 ± 0.01	0.0650	0.2086	0.26

NOTE: Values are means ± SD ($n = 100$ cells per cell line). *z*-score is for the cancer cell line. Raw *P* values are derived from univariate two-group *t* tests. StepBoot *P* values reflect a step-down adjustment for multiple comparisons and do not rely on the assumption of normality (29). See http://cpl.colostate.edu/morph for details.

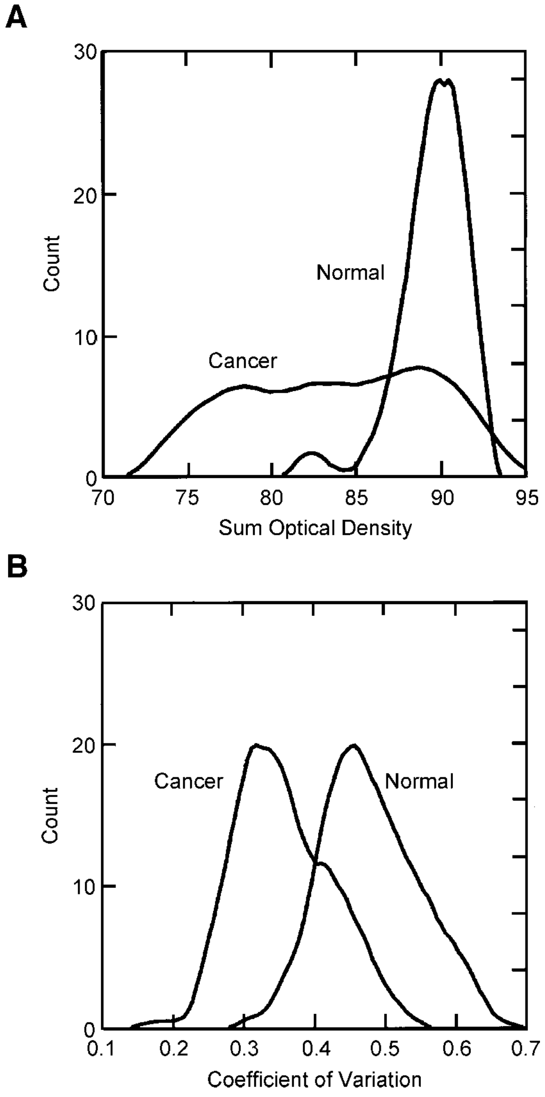

Figure 2. Probability density plots (Epanechnikov KD smoother) of sum absorbance (sum optical density) (**A**) and coefficient of variation (**B**) by cell line.

each parameter. Only 22 (11%) observations, 15 cells from the cancer cell line and 7 cells from the normal cell line, were misclassified (see Table 3). The canonical coefficients from the model, which are used to classify new observations, may also be used to construct a weighed mean, or summary measure for the z-scores, which Bacus et al. called NG. In this data set, NG = [0.866(coefficient of variation) + 0.295(PgDNA) − 0.401(sum variance) + 0.253(cell Feret Y) − 0.789(slope) + 0.225(valley)] / 6. NG can also be interpreted as a classifier, d(**x**), where the threshold for our experiment is −0.24. Density plots of NG by cell line are shown in Fig. 3a. Observations with values that fall into the area under both curves will be misclassified if they are in the normal cell line and have NG < −0.24 (n = 4) or in the cancer cell line and have NG > −0.24 (n = 16). The mean NG in the cancer cell line is −0.47 ± 0.20 and mean NG in the normal cells is 0.00 ± 0.12. This method rests entirely on the performance of the underlying LDA and therefore does not appear in the comparisons in Table 3. A small improvement can be made to the misclassification rate based on the information in the ROC curve; assigning observations to the cancer cell line when Pr(Y = 1) > 0.28 gave 90% sensitivity and 92% specificity while reducing the misclassification rate by about 1%.

Stepwise LR performed better than LDA. The misclassification rate was 10.5%. The predicted probabilities for membership in the cancer cell line were saved and plots by cell line are shown in Fig. 3b. The stepwise procedure applied to raw data selected four parameters for the discriminant function: valley, coefficient of variation, diagonal moment, and sum absorbance. We ran the model again with z-scores as input; the parameters selected were valley, coefficient of variation, diagonal moment, and PgDNA. The correlation between PgDNA and sum

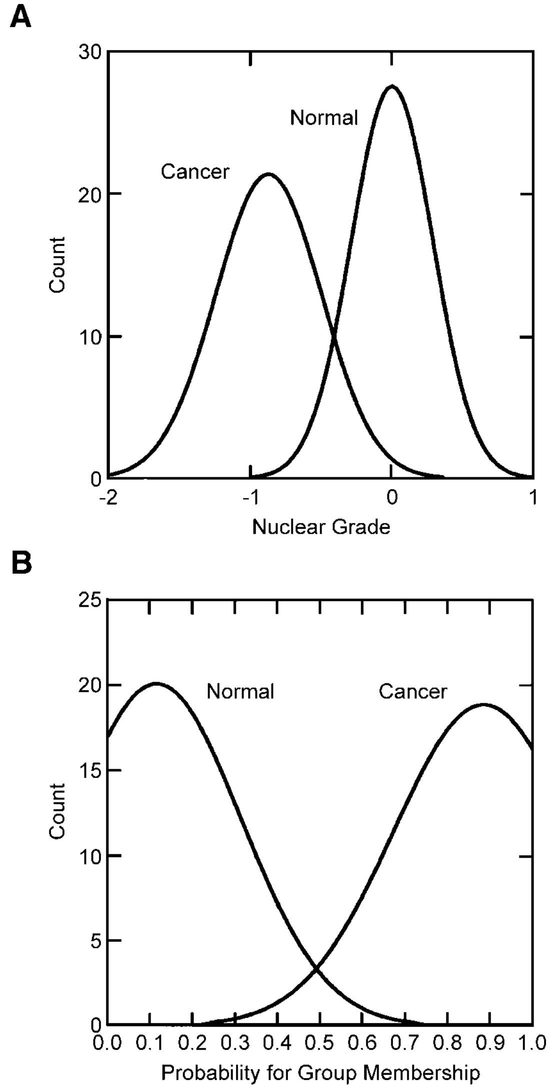

Figure 3. Probability density plots (normal density function) of NG derived from z-scores and discriminant analysis (**A**) and probabilities for class membership from LR (**B**).

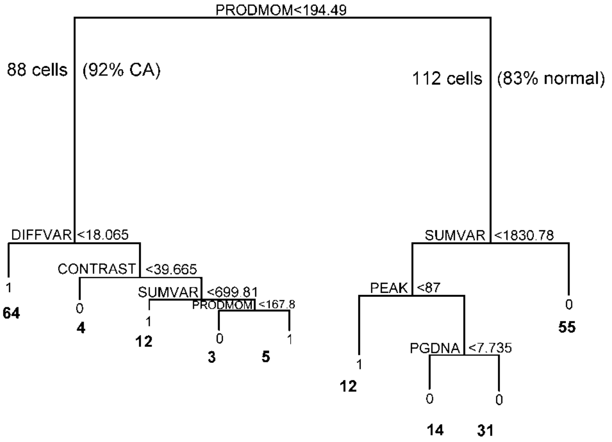

Figure 4. Classification tree produced by CART using six morphometric parameters and 200 observations. Leaves (terminal nodes) are classified 1 for cancer and 0 for normal. The length of the branches is proportional to the importance of the split. The *number of cells at each leaf* is below the classification in *bold*; only one leaf, the *third from the right*, with n = 14 cells, has both classes represented (the split is 50/50).

absorbance is 0.999. The resubstitution error rate was not different, so we did not estimate the misclassification rate both ways. The mean predicted probability for membership in the normal cell line is 0.14 ± 0.266 (n = 9 errors) and mean predicted probability for membership in the cancer cell line is 0.86 ± 0.235 (n = 12 errors) using the default threshold of 0.5. A marginally lower misclassification rate can be found by examining the ROC curve; assigning observations to the cancer cell line when $\Pr(Y = 1) > .36$ gave 92% sensitivity and 88% specificity while reducing the misclassification rate by about 0.5%.

Parameter selection in both LDA and LR was dependent on the data and the choice of classifier; different subsets of parameters were selected for each subset of data, $L - L_v$, in the cross-validation models. Only PgDNA appeared in all 10 partitions for discriminant analysis, for example, and in 8 of 10 models for LR. This is due to the strong correlations among the parameters. Backward selection, the preferred method because variables behave differently in concert than singly, does not work well with these data, again due to the strong correlations. In the cross-validation macro for LR for example, 10 to 20 variables were selected depending on the partition, and the misclassification rate was high due to overfitting. Table 3 shows the set of parameters for each method based on stepwise selection on the full 200 observations.

CART, technically known as binary recursive partitioning, is widely used in the behavioral sciences and has recently been applied to the high-dimensional data of proteomic assays (30). The misclassification was the same as that for LDA after adjusting the default parameters for stopping. We present the graphic results for the process with a tree grown on L and pruned based on the minimum error criterion (see Fig. 4). The importance of each element of **x** is evident in the vertical distance between nodes. Interactions among parameters can be found by observing which parameters split the data along the left and right branches of the tree. For example, PgDNA and peak are predictors only when product moment is >194, while difference variance is a predictor only when product moment is <194.

KNN and KD do not associate P values with parameters, so there is no option for stepwise elimination of variables that do not contribute to correct classification. We computed cross-validation misclassification error rates based on all 39 parameters for both methods and compared them with the misclassification rates based on the 32 parameters that have $P < 0.20$ using the bootstrap (Table 4). There was no improvement for KNN, but the error rate for the KD method was 2% lower with the extraneous features removed. There is no reason to expect KD to perform well with the same subset of features selected by LR or LDA, and the misclassification rate was the same applying the method to 32 features or to the four selected by LR on the full data set. Table 3 shows the error rates for both methods applied to 32 features.

Because the area under the ROC curve was similar for LR, LDA, and KD, we examined the ROC data to find a cut point that would optimize the misclassification rate for KD. Setting the cut point at 0.42 gives 87% specificity and 91% sensitivity and reduced the misclassification rate by 2%. The ROC curves for these three methods, which had the greatest area under the ROC curve, are shown in Fig. 5.

Other Methods. We also tried quadratic discriminant analysis and experimented with neural networks and the support vector machine implemented in nnet() and svm() in R; the results (not shown) were poor. The machine learning techniques are better suited for the high-dimensional data encountered in genomic and proteomic research (18), where there are several thousand features available. We also applied the CART, logistic, and neural network techniques implemented in iMiner to L. These results are available on our Web site (http://cpl.colostate. edu/morph).

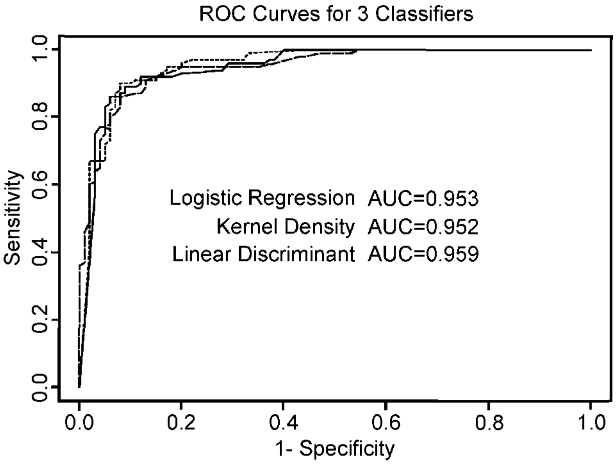

Figure 5. ROC curves for the three methods with the greatest area under the ROC curve. *Solid line,* ROC curve for LR; *short dashes,* LDA; *longer dashes,* KD method.

Discussion

The goal of this study was to identify classification methods that are best suited for the evaluation of nuclear morphometric data. To do this, a very simple cell culture experiment was designed to test the ability of various methods to correctly classify cells into the populations from which they were obtained. The main result is that LR has the lowest misclassification rate using default settings, consistent with theory (15), empirical work (31), and simulation studies (17); its superiority over LDA when the data are not multivariate normal is well established. None of the methods we applied achieved complete separation of the data into cancer cell line and normal cell line for all the validation samples. Correct classification was generally higher for the normal cell line (*i.e.*, specificity was greater than sensitivity). CART provided the most information on the structure of the data and potential interactions among parameters but has difficulty in capturing a simple additive structure. The following discussion addresses factors that should be considered in the presentation and analysis of morphometry data.

Standardization. The benefit of standardizing data prior to analysis is less obvious in our example than it is in theory. The transformation of centering alone improves the numerical properties of the data, and further division by the SD of the referent group, the calculation of the z-score, provides an intuitively appealing measure in SD units. The improvement in numerical stability achieved by this process would be more apparent in a larger data set. Some applications, KNN, for example, require standardization of the inputs to avoid the problem of dissimilarities of scale. The drawback to the procedure we used, standardizing with respect to a particular reference group, is that while most statistical software has an option to produce z-scores, they will be column specific. That is, for each element $\mathbf{x_i}$, the transformation is done with respect to the mean and SD for the parameter over all classes, so that some programming is required to accomplish the transformation where the mean and SD are restricted to a particular subset of the data. The performance of the classifiers was not affected by the choice of transformation in our data set.

Multiple Comparisons. Typically, analyses in the morphometry literature ignore the problem of multiple endpoints as well as the distribution of the morphometric measures. Adjustment for multiple comparisons in a nonparametric procedure leads to more conservative conclusions than those we would have drawn by examining the P values on a list of 39 t tests. The step-down bootstrap is a better choice than the traditional Bonferroni adjustment on theoretical grounds; the advantages increase dramatically as the number of tests increases, divergence from the normal distribution increases, and in the presence of correlation among the test statistics. The feature-by-feature analysis presented in Table 4 is particularly useful in situations where it is not known which parameters might be important (32) and should be considered with appropriate adjustment for multiple comparisons (column 6).

Methods of variable selection that use information on the known classes perform multiple comparisons iteratively; the resulting models, if not adjusted for shrinkage, are characterized by P values that are too small and by biased parameter estimates. The stepwise selection we used for LR and LDA would lead to highly suspect parameters if the objective were model building and parameter estimation rather than predicting class membership.

Data Reduction. The morphometry data are inherently redundant and multivariate in nature. A single summary measure, data reduction techniques, or computer methods that select only the relevant information hold an obvious appeal. Bacus et al. (10) presented a summary measure, histologic grade. Its construction was the result of an iterative, labor-intensive process in which more than 100 possible combinations of features were examined. Among the characteristics of the resulting measure were "a high histologic grade at the end of the neoplastic process, . . . a monotonically increasing histologic grade from 0 to the endpoint over the period of preinvasive neoplastic growth, [and] a generally Gaussian distribution of the histologic grade for the distribution of samples taken from . . . normal tissues." They were combining data reduction and variable selection. The same authors proposed a similar but more objective summary measure in 1999 (22): they used LDA to select the important features from a collection of features standardized against the corresponding features from normal tissue and combined them into a weighed mean, where the weights are the canonical coefficients estimated for the discriminant function. The resulting NG inherits its predictive strength from the underlying linear discriminant function, however, and the probabilities for class membership estimated from LR can be used in the same way with more reliable results.

Data reduction, as opposed to variable selection, does not use information on the outcome or the known classes. The best approach would be to eliminate unimportant variable from the analysis based on prior knowledge or other criteria not requiring reference to the known classes. This approach has limited usefulness for the morphometry data because many of the features have no apparent clinical interpretation. Statistical approaches, such as variable clustering and nonlinear generalizations of principal components, have not been applied to these data to our knowledge.

Classification. LR has not been widely applied to morphometric data. A few exceptions are Kavantzas et al. (33), Veltri et al. (6), and Acker et al. (34). LR performs the same task as LDA, but the discrimination is conditional on the observed values of the random vectors in the training set. Efron (15) shows that while LR is less efficient (estimates will have greater variance) than LDA when the data are multivariate normal, it is robust to departures from normality. When the assumption of multivariate normal is violated, parameter estimates in the normal discriminant procedure are biased away from zero when the true parameter is not zero, which means that the tendency will be to overestimate the importance of the underlying variable. The estimated probabilities for membership in the *event* category, defined in our model to be the cancer cell line, can be interpreted in the same way we interpret the NG. Figure 3b shows probability density plots of the estimated probability of class membership from the regression of cell type on valley, coefficient of variation, diagonal moment, and sum absorbance. The misclassification rate with the cut point set at 0.5 argues in favor of the use of LR over the other methods we tested. The departures from multinormal distribution in the morphometric data are nontrivial, which suggests that the information about the importance of each parameter used in the classifier will be more reliably assessed by LR in cases where information about the contributing parameters is of interest. A limitation of LR is the requirement that any interactions to be tested must be specified by the analyst. Another is that the predictors should be linear in the logit, which can often be accomplished with an appropriate transformation.

An alternative measure of classification accuracy, the area under the ROC curve, groups LDA, LR, and KD more closely than the misclassification rate. The differences are very small, and given the SD of each, it is expected that the order would shift with different data sets. Optimizing the misclassification rate by adjusting the cut point produced a lower misclassification rate for each method, with the greatest reductions for LDA and KD, giving LDA the lowest misclassification rate (9%). Clearly, in settings where there is particular interest in sensitivity or specificity, adjusting the cut point based on exploration of the ROC data will be useful.

We found only one example of CART models applied to morphometric data (35). The misclassification rate for the S-Plus version of CART was not markedly different from LR and the same as LDA in SAS. CART has two distinct advantages over LR: the classifier is not assumed to be linear in **x**, and it is not necessary for the analyst to specify which interactions are to be tested, assuming the model is not simply additive. In a data set with 39 predictors, there are 741 possible two-way interactions and more than 10,000 three-way interactions. CART analysis can be used to complement LR; Nelson et al. (36) argue for the use of the two methods in concert: CART for elucidating the important subgroups and LR for estimating odd ratios in the subgroups. We examined the performance of LR with the variables and implied interactions from the CART model shown in Fig. 4; the interaction terms were not significant. As in LR, the classification variable can be continuous, categorical, or binary in CART; there are no restrictions on or assumptions about the distribution of the variables to be used in the classification process, nor is there any difficulty in handling complex, high-dimensional data sets containing mixed data types. In our data, the misclassification rate for new observations was slightly higher with CART than with LR, but studies in other fields have shown the performance of CART models to improve with larger training sets and more complex structure, especially if bagging or boosting can be used to improve their accuracy (18).

Limitations. Complete separation at the cell level is not possible in these data irrespective of the classification method used. A review of published articles in the morphometric literature reveals similar findings for *in vivo* data, across diverse model systems, and for different sampling procedures. See, for example, Markopoulos et al. (13), Pantazopoulos et al. (14), Boone et al. (9, 38), Bacus (37), and Poulin et al. (39). These results are central to the problem of determining the applications in which morphometry data are likely to be useful. Consider the difference between the problem of determining whether two populations are distributed differently and the problem of assigning to one population a single observation drawn at random from a mixture of the two populations based on a single characteristic. For the sake of argument, suppose the characteristic for the two populations is normally distributed and differs only in its mean (*i.e.*, the variance is the same in both populations). If the distributions overlap, as they do in Figs. 2 and 3, the assignment of a new case selected at random will have a probability for misclassification that is a function of the amount of overlap; without information on other characteristics of the population, 100% correct classification is not possible. Conversely, if we want to estimate a confidence interval around either mean, the width of that interval can be made as narrow as we please by increasing the number of observations drawn from the mixture of the two populations. If the mean of a parameter (or combination of parameters) tends to shift in the presence of, say, a chemopreventive agent administered to a subject over time, then the hypothesis that a competing agent has the same effect can be accepted (or rejected) at any size test and power of the test that the investigator is comfortable with, simply by selecting an adequate number of subjects at the beginning of the experiment. The only way to improve the misclassification rate for new observations, however, is to identify another clinical characteristic that has discriminatory power. An example of this is given in Veltri et al. (40), where biochemical recurrence of prostate cancer survivors was equally well predicted by quantitative NG or the Gleason score. When quantitative NG and the Gleason score were combined, sensitivity, specificity, and accuracy increased.

The stepwise methods we used select four or five parameters, depending on where the threshold for entry into the model is set. A different subset of features emerged for each of the 10 partitions used for cross-validation, illustrating the point that variable selection is made arbitrary by collinearity (41). Models for classification in a variety of tissue or cancer types are characterized by even more striking differences and may have only one or two features in common (10). Given the redundancy in the data, it is not surprising but does imply that, for each new application, an extensive feature-finding trial should be performed, analogous to dose finding in a new pharmaceutical agent.

Our data set was carefully selected to maximize the signal. With less careful cell selection, noise levels will be higher and may introduce bias. How the various classifiers perform under varying conditions has been much studied, and if one message is clear, it is that no one method is universally better than the others. The results of several methods we tested have been reported for nuclear morphometry data on a variety of tissue types, and many studies report misclassification rates in the 10% to 15% range, not markedly higher than the rates we achieved in an ideal setting with optimal thresholds. The differences between classification of cells and whole tissue samples may be small. Poulin et al. (39), for example, found little difference in classification accuracy between single cell analysis and histometric analysis.

Conclusions. The value of the classifiers that produce an estimated probability for group membership, which can be used to establish a threshold for group assignment, is apparent in their potential for interpretation as surrogate endpoint biomarkers or risk scores. The examples in Fig. 3 establish the extremes for disease *versus* no disease for a particular function. The function, or algorithm, could be applied to similarly acquired populations of cells from subjects at various levels of known risk or disease progression. We expect that the number of cells with characteristics similar to the most extreme disease class will increase with risk or disease progression, causing the function to shift toward the disease distribution. This has been done with NG (22) but not, to our knowledge, with the LR classifier, which has a lower misclassification rate at the default threshold. Predictions from a good CART model or KD could also be used in this way. Further research is needed to extend the methods comparison with *in vivo* data, larger data sets, and other cancers; the clustering of methods based on the area under the ROC curve strongly suggests the need for exploration with each new tissue or model. Interindividual variability, which was not a factor in this experiment, is an important component of *in vivo* data. For many applications, the prevalence of a condition in the population is relevant, and appropriate prior probabilities can be assigned based on expected event rates. In our data, adjusting the prior probability for the cancer cell line to 0.05 in LDA increased the misclassification rate to 16.5%. The development of risk indices and their statistical properties is a necessary step toward incorporating morphometric features into chemoprevention and risk reduction trials.

Acknowledgments

We thank Anna Barón and Diane Fairclough for thoughtful comments and suggestions on an earlier version of this manuscript and the anonymous reviewers whose diligence was evident in the thoroughness of their remarks; in responding to their comments, we hope to have improved the clarity of the presentation.

References

1. Baak JP. The principles and advances of quantitative pathology. Anal Quant Cytol Histol 1987;9:89-95.
2. Carr I, Pettigrew N. How malignant is malignant? A brief review of the microscopic assessment of human neoplasms, and the prediction of whether they will metastasize and kill. Clin Exp Metastasis 1991; 9:127-37.
3. Collan Y, Torkkeli T, Pesonen E, Jantunen E, Kosma VM. Application of morphometry in tumor pathology. Anal Quant Cytol Histol 1987; 9:79-88.
4. Gil J, Wu H, Wang BY. Image analysis and morphometry in the diagnosis of breast cancer. Microsc Res Tech 2002;59:109-18.
5. Millot C, Dufer J. Clinical applications of image cytometry to human tumor analysis. Histol Histopathol 2000;15:1185-200.
6. Veltri RW, Partin AW, Miller MC. Quantitative nuclear grade (QNG): a new image analysis-based biomarker of clinically relevant nuclear structure alterations. J Cell Biochem Suppl 2000;35:151-7.
7. Doudkine A, MacAulay C, Poulin N, Palcic B. Nuclear texture measurements in image cytometry. Pathologica 1995;87:286-99.
8. Hamilton PW, Bartels PH, Wilson RH, Sloan JM. Nuclear texture measurements in normal colorectal glands. Anal Quant Cytol Histol 1995;17:397-405.
9. Boone CW, Stoner GD, Bacus JV, et al. Quantitative grading of rat esophageal carcinogenesis using computer-assisted image tile analysis. Cancer Epidemiol Biomarkers & Prev 2000;9:495-500.
10. Bacus JW, Bacus JV, Stoner GD, Moore GW, Kelloff GJ, Boone CW. Quantitation of preinvasive neoplastic progression in animal models of chemical carcinogenesis. J Cell Biochem Suppl 1997;29:21-38.
11. Poulin N, Boiko IV, MacAulay C, et al. Nuclear morphometry as an intermediate endpoint biomarker in chemoprevention of cervical carcinoma using α-difluoromethylornithine. Cytometry (Commun Clin Cytom) 1999;38:214-23.
12. Palcic B. Nuclear texture: can it be used as a surrogate endpoint biomarker? J Cell Biochem Suppl 1994;19:40-6.
13. Markopoulos C, Karakitsos P, Botsoli-Stergiou E, et al. Application of the learning vector quantizer to the classification of breast lesions. Anal Quant Cytol Histol 1997;19:453-60.
14. Pantazopoulos D, Karakitsos P, Iokim-Liossi A, Pouliakis A, Botsoli-Stergiou E, Dimopoulos C. Back propagation neural network in the discrimination of benign from malignant lower urinary tract lesions. J Urol 1998;159:1619-23.
15. Efron B. The efficiency of logistic regression compared to normal discriminant analysis. J Am Stat Assoc 1975;70:892-8.
16. Bull SB, Donner A. The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. J Am Stat Assoc 1987;82:1118-22.
17. Barón AE. Misclassification among methods used for multiple group discrimination—the effects of distributional properties. Stat Med 1991;10:757-66.
18. Dudoit S, Fridlyand J, Speed T. Comparison of discrimination methods for the classification of tumors using gene expression data. J Am Stat Assoc 2002;97:77-87.
19. Hand DJ. Discrimination and classification. Chichester: Wiley & Sons; 1981.
20. Hastie T, Tibshirani RJ, Freidman J. The elements of statistical learning: data mining, inference, and prediction. New York: Springer-Verlag; 2001.
21. Gurley AM, Hidvegi DF, Bacus JW, Bacus SS. Comparison of the Papanicolaou and Fulgen staining methods for DNA quantification by image analysis. Cytometry 1990;11:468-74.
22. Bacus JW, Boone CW, Bacus JV, et al. Image morphometric nuclear grading of intraepithelial neoplastic lesions with applications to cancer chemoprevention trials. Cancer Epidemiol Biomarkers & Prev 1999;8:1087-94.
23. Morrison DF. Multivariate statistical methods. New York: McGraw-Hill Publishing Co.; 1990.
24. Hosmer DW, Lemeshow S. Applied logistic regression. New York: John Wiley & Sons, Inc.; 2000.
25. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. New York: Chapman & Hall/CRC; 1998.
26. Pantazopoulos D, Karakitsos P, Iokim-Liossi A, Pouliakis A, Dimopoulos K. Comparing neural networks in the discrimination of benign from malignant lower urinary tract lesions. Br J Urol 1998;81: 574-9.
27. Pantazopoulos D, Karakitsos P, Pouliakis A, Iokim-Liossi A, Dimopoulos MA. Static cytometry and neural networks in the discrimination of lower urinary system lesions. Urology 1998;51:946-50.
28. Westfall PH, Young SS. Resampling-based multiple testing. New York: John Wiley & Sons, Inc.; 1993.
29. Westfall PH, Tobias RD, Rom D, Wolfinger RD, Hochberg Y. Multiple Comparisons and multiple tests using the SAS system. Cary, North Carolina: SAS Institute Inc.; 1999.
30. Qu Y, Adam B-L, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. Clin Chem 2002;48:1835-43.
31. Marshall G, Grover FL, Henderson WG, Hammermeister KE. Assessment of predictive models for binary outcomes: an empirical approach using operative death from cardiac surgery. Stat Med 1994;133:1501-11.
32. McGinley JN, Knott KK, Spoelstra NS, Thompson HJ. A Quantitative assessment of the effect of storage temperature on the nuclear morphometry of cells obtained from sputum. Charlotte, North Carolina: National Society for Histotechnology Symposium; 2001.
33. Kavantzas N, Lazaris AC, Chatzigianni E, Davaris PS. The nuclear morphometry by image analysis in the histopathologic diagnosis of lung cancer. J Exp Clin Cancer Res 2000;19:201-6.
34. Acker SM, Nicholson JH, Rust PF, Maize JC. Morphometric discrimination of melanoma *in situ* of sun-damaged skin from chronically sun-damaged skin. J Am Acad Dermatol 1998;39:239-45.
35. Thiele J, Kvasnicka HM, Zirbes TK, et al. Impact of clinical and morphological variables in classification and regression tree-based survival (CART) analysis of CML with special emphasis on dynamic features. Eur J Haematol 1998;60:35-46.
36. Nelson LM, Bloch DA, Longstreth WT, Shi H. Recursive partitioning

for the identification of disease risk subgroups: a case-control study of subarachnoid hemorrhage. J Clin Epidemiol 1998;51:199-209.

37. Bacus JW. Cervical cell recognition and morphometric grading by image analysis. J Cell Biochem Suppl 1995;23:33-42.

38. Boone CW, Bacus JW, Bacus JV, Steele VE, Kelloff GJ. Properties of intraepithelial neoplasia relevant to the development of cancer chemopreventive agents. J Cell Biochem Suppl 1997;28-29:1-20.

39. Poulin N, Susnik B, Guillaud M, Doudkine A, Worth A, Palcic B. Histometric texture analysis of DNA in thin sections from breast biopsies. Application to the detection of malignancy-associated changes in carcinoma *in situ.* Anal Quant Cytol Histol 1995;17:291-9.

40. Veltri RW, Miller MC, Partin AW, Coffey DS, Epstein JI. Ability to predict biochemical progression using Gleason score and a computer-generated quantitative nuclear grade derived from cancer cell nuclei. Urology 1996;48:685-91.

41. Harrell FE. Regression modeling strategies. New York: Springer-Verlag; 2001.