Journal of Biomedicine and Biotechnology • 2005:2 (2005) 147–154 • DOI: 10.1155/JBB.2005.147

RESEARCH ARTICLE

Classification and Selection of Biomarkers in Genomic Data Using LASSO

Debashis Ghosh[1] and Arul M. Chinnaiyan[2]

[1]Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA
[2]Departments of Pathology and Urology, University of Michigan, 1300 Catherine Road, Ann Arbor, MI 48109-1063, USA

Received 3 June 2004; accepted 13 August 2004

High-throughput gene expression technologies such as microarrays have been utilized in a variety of scientific applications. Most of the work has been done on assessing univariate associations between gene expression profiles with clinical outcome (variable selection) or on developing classification procedures with gene expression data (supervised learning). We consider a hybrid variable selection/classification approach that is based on linear combinations of the gene expression profiles that maximize an accuracy measure summarized using the receiver operating characteristic curve. Under a specific probability model, this leads to the consideration of linear discriminant functions. We incorporate an automated variable selection approach using LASSO. An equivalence between LASSO estimation with support vector machines allows for model fitting using standard software. We apply the proposed method to simulated data as well as data from a recently published prostate cancer study.

INTRODUCTION

DNA microarrays simultaneously gauge the expression of thousands of genes in clinical samples. In this paper, we focus on cancer studies, where gene expression technologies have been applied extensively (Alizadeh et al [1]; Khan et al [2]; Dhanasekaran et al [3]). Obtaining large-scale gene expression profiles of tumors should theoretically allow for the identification of subsets of genes that function as prognostic disease markers or biologic predictors of therapeutic response. Because the data are highly multivariate and complex, it is important to develop automated statistical methods to detect systematic signals in gene expression patterns.

In cancer studies, analyses have typically focused on one of three problems. First, investigators have looked for genes that discriminate neoplastic from benign tissue. Statistically, this is the problem assessing differential expression of genes and has been studied by several authors; see, for example, Efron et al [4]. A second problem is clustering the samples to find subtypes of disease using algorithms such as those in [5]. The final class of problems is classification or supervised learning, which involves using the profile to predict some clinical outcome, such as the stage of disease. Suppose that in this instance, we treat the gene expression profile as the independent variables and tissue type as the response. A particular feature of microarray experiments is that the dimension of the predictor space (number of genes) is typically larger than the number of samples. This is known as the "large p, small n" paradigm (West [6]), so classification methods must take this into account.

One method to do this is apply prefiltering criteria in which the candidate number of genes for building a classifier is smaller than the number of samples. For example, Dudoit et al [7] performed a systematic comparison of several discrimination methods for the classification of tumors based on microarray experiments. However, they must perform an initial reduction in the number of predictors before building the classifier.

We wish to consider the joint effects of genes in determining classification rules for discriminating tumors. There are two assumptions that drive our proposed methodology. First, we assume that the joint effects of multiple genes must be considered in discriminating classes of disease. Recently, much attention has been given to the finding that a 70-gene signature can predict breast cancer survival (van't Veer et al [8]; van de Vijver et al [9]). However, most such gene signatures have been constructed using univariate methods. It seems reasonable to consider joint models, as genes are correlated because of their mutual involvement in disease pathways.

Correspondence and reprint requests to D. Ghosh, Department of Biostatistics, University of Michigan, 1420 Washington Heights, Ann Arbor, MI 48109-2029, USA, E-mail: ghoshd@umich.edu

The second assumption is that there are individual genes that can discriminate classes. This is different from the latent factor and partial least squares proposals put forth by other authors (West [6]; Nguyen and Rocke [10]), where linear combinations of all available genes are used to predict the outcome. We seek to develop interpretable models for classification; for this purpose, using individual genes for predictors rather than linear combinations of genes seems reasonable.

In this paper, we develop classification rules based on the consideration of measures of diagnostic accuracy. In particular, we are interested in finding gene expression profiles that can discriminate between two populations. A unique challenge is posed because of the large p, small n problem. Our solution is to combine the problems of variable selection and classification. We suggest an approach for classification using the LASSO approach (Tibshirani [11]). An advantage of this approach is that some of the effects of the variables in these models are estimated to be exactly zero. These will represent genes that have no discriminatory power between the two classes, while those with nonzero coefficients will represent genes that can separate classes of tumors successfully. Thus, a byproduct of the approach is the generation of a gene list. We exploit an equivalence between LASSO and support vector machines (SVMs) in order to fit the proposed classifier. The structure of the paper is as follows. In "materials and methods," we provide background on the data structures observed and the motivation based on biomarker combinations, which leads to the use of linear discriminant functions. We also provide a review of LASSO estimation (Tibshirani [11]) in this section. The latter two techniques are then involved in the proposed estimation procedure, described in "results and discussion." There, we also describe how to implement the proposed method using software for SVMs. Issues of model selection are also discussed. We describe the application of the proposed methodologies to simulated data and data from a recent cancer profiling study (Dhanasekaran et al [3]) in "prostate cancer gene expression data." Finally, some concluding remarks are made in "conclusion."

MATERIALS AND METHODS

Let $\mathbf{a}^T$ denote the transpose of the vector $\mathbf{a}$. For the ith sample ($i = 1, \dots, n$), we let $\mathbf{X}_i = [X_{i1} \cdots X_{ip}]^T$ denote the $p \times 1$ gene expression profile vector (ie, X_{ij} is the gene expression measurement of the jth gene, $j = 1, \dots, p$). We suppose that the data have already been preprocessed and normalized. In addition, it is assumed that the gene expression data are standardized so that for each gene, the mean is zero and standard deviation one. Let g_i denote the tumor class for the ith sample ($i = 1, \dots, n$); we assume that there are two classes so that g_i takes values $g \in \{0, 1\}$. Here and in the sequel, we will refer to $g = 1$ as the diseased class and $g = 0$ as the healthy class; however, the methods proposed here are applicable to any two-class setting. In "LASSO estimation," we assume the existence of a continuous response variable Y_i for the ith sample ($i = 1, \dots, n$).

ROC curves and optimal biomarker combinations

Our approach is to consider each measurement from a microarray for a single gene as a diagnostic test. Thus, for each subject, we have a high-dimensional vector of diagnostic test results. We then want to utilize this information in a way to separate the two populations of patients. This issue of finding combinations of biomarkers to accurately classify patients has been considered by Su and Liu [12], Baker [13], and Pepe and Thompson [14] in the statistical literature.

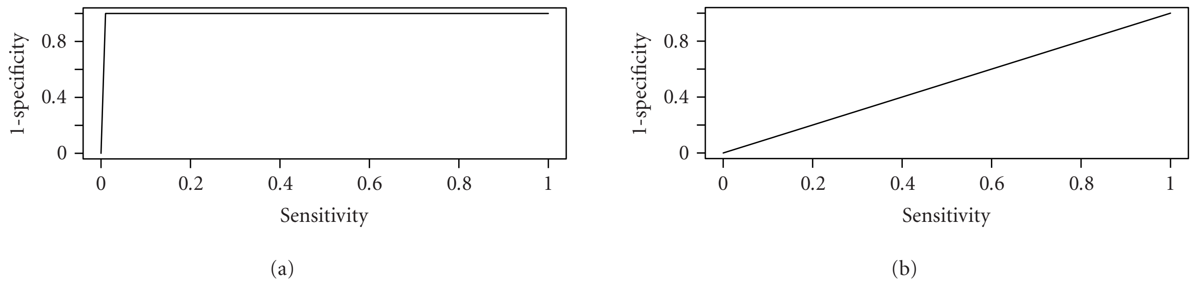
To combine information across the high-dimensional vector of gene expression profiles, we consider linear combinations of the form $\beta_0^T \mathbf{X}_i$, $i = 1, \dots, n$. Without loss of generality, we will also assume that larger values of this linear combination corresponding to increasing likelihood of having $g = 1$. While the method can be easily extended to incorporate interactions between gene expression measurements, we focus on consideration of the main effects for purposes of exposition.

Suppose $\mathbf{X}^D$ represents the gene expression profile for a typical cancer specimen (ie, $g = 1$), and $\mathbf{X}^{\bar{D}}$ is the corresponding profile for a randomly chosen benign specimen. Note that in our situation, the diagnostic test is the linear combination $\beta_0^T \mathbf{X}$. One relevant quantity is the false positive rate based on a cutoff c, defined to be $FP(c) = P(\beta_0^T \mathbf{X} > c | g = 0)$. Similarly, the true positive rate is $TP(c) = P(\beta_0^T \mathbf{X} > c | g = 1)$. The true and false positive rates can be summarized by the receiver operating characteristic (ROC) curve, which is a graphical presentation of $\{FP(c), TP(c) : -\infty < c < \infty\}$. The ROC curve shows the tradeoff between increasing true and false positive rates. Tests that are have $\{FP(c), TP(c)\}$ values close to $(0, 1)$ indicate perfect discriminators, while those with $\{FP(c), TP(c)\}$ values close to the 45° line in the $(0, 1) \times (0, 1)$ plane are tests that are unable to discriminate between the diseased and healthy populations. Examples of ideal and noninformative ROC curves are given in Figures 1a and 1b.

While the specificity and sensitivity of a diagnostic test depend on the cutoff value chosen, a useful summary measure to consider is the area under the ROC curve. It can be shown mathematically that the area under curve is $P(\beta_0^T \mathbf{X}^D > \beta_0^T \mathbf{X}^{\bar{D}})$ (Bamber [15]). Under a binormal probability model, Su and Liu [12] showed that this quantity is optimized using the linear discriminant function. This motivates our choice of consideration of these variables. We next present an algorithm for estimation of these functions.

Linear discriminant functions by optimal scoring

While linear discriminant analysis (LDA) is typically calculated using matrix algebra techniques, an alternative method of calculating them is through the use of optimal scoring (Hastie et al [16, 17]). In this method, the



(a)

(b)

Figure 1. Receiver operating characteristic (ROC) curves for (a) ideal and (b) noninformative tests.

problem of classification into two groups is reexpressed as a regression problem based on quantities known as optimal scores.

The point of optimal scoring is to turn the categorical class labels into quantitative variables. Let $\boldsymbol{\theta}(g) = [\theta(g_1),\ldots,\theta(g_n)]^T$ be the $n \times 1$ vector of quantitative scores assigned to $\mathbf{g}$ for the kth class. The optimal scoring problem involves finding the vector of coefficients $\boldsymbol{\eta} \equiv (\eta_1, \eta_2, \ldots, \eta_p)$ and the scoring map $\theta : \{0,1\} \rightarrow R$ that minimize the following average squared residual:

$$\text{ASR} = n^{-1}\sum_{i=1}^{n}\{\theta(g_i) - \mathbf{X}_i^T\boldsymbol{\eta}\}^2. \tag{1}$$

Let $\mathbf{Z}$ be an $n \times 2$ matrix with the ith row equal to $(1,0)$ if $g_i = 1$ and $(0,1)$ if $g_i = 0$ $(i = 1,\ldots,n)$. The optimal scores are assumed to be mutually orthogonal and normalized with respect to an inner product. Thus, the minimization of (1) is subject to the constraint $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\|^2 = 1$, where $\boldsymbol{\Theta} = [\theta(0)\ \ \theta(1)]^T$ is a 2×1 vector of the optimal scores. Hastie et al [16] state that the minimization of this constrained optimization problem leads to estimates of $\boldsymbol{\eta}$ that are proportional to the discriminant variables (ie, the discriminant function) in LDA. In particular, they propose the following algorithm for the estimation of the LDA functions

(1) Choose an initial score matrix $\boldsymbol{\Theta}_0$ satisfying $\boldsymbol{\Theta}_0^T\mathbf{D}_p\boldsymbol{\Theta}_0 = \mathbf{I}$, where $\mathbf{D}_p = \mathbf{Z}^T\mathbf{Z}/n$. Let $\boldsymbol{\Theta}_0^* = \mathbf{Z}\boldsymbol{\Theta}_0$.

(2) Let $\mathbf{X}$ be the $n \times p$ matrix with ith row $\mathbf{X}_i$. Fit a linear regression model of $\boldsymbol{\Theta}_0^*$ on $\mathbf{X}$, yielding fitted values $\hat{\boldsymbol{\Theta}}$. Let $\hat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.

(3) Obtain the eigenvector matrix $\boldsymbol{\Phi}$ of $\boldsymbol{\Theta}_0^{*T}\hat{\boldsymbol{\Theta}}$; the optimal scores are then $\boldsymbol{\Theta}^* = \boldsymbol{\Theta}_0\boldsymbol{\Phi}$.

(4) Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \boldsymbol{\Phi}^T\hat{\mathbf{f}}(\mathbf{x})$.

As mentioned before, a problem with attempting to apply standard linear discriminant function methods to the data here is that there is not a numerically unique solution because p is larger than n. Thus, some type of regularization is needed. Our approach is based on the LASSO, which is described in the next section.

LASSO estimation

We suppose that our data are $(Y_i, \mathbf{X}_i)$, where Y_i $(i = 1,\ldots,n)$ is a continuous variable. The LASSO solution is to the optimization problem of minimizing

$$\sum_{i=1}^{n}(Y_i - \beta^T\mathbf{X}_i)^2 + \lambda\sum_{j=1}^{p}|\beta_j|, \tag{2}$$

where $\beta = (\beta_1,\ldots,\beta_p)$ and $\lambda \geq 0$ is a penalty term. Thus, the constraint that is utilized is an L_1 constraint. An alternative way of formulating (2) is to minimize $\sum_{i=1}^{n}(Y_i - \beta^T\mathbf{X}_i)^2$, subject to the constraint that $\sum_{j=1}^{p}|\beta_j| \leq t$. Note that in the absence of the constraint, the solution is given by the ordinary least squares (OLS) estimator. If the usual OLS estimator satisfies the constraint, then the LASSO and OLS estimates of β coincide. However, for smaller values of t, some of the components of β are estimated to be zero. In the linear regression setting, LASSO estimation has been considered by Tibshirani [11].

For a given value of t, minimization of $\sum_{i=1}^{n}(Y_i - \beta^T\mathbf{X}_i)^2$ subject to an L_1 constraint on the components of β is a quadratic programming problem with 2^p linear equality constraints. A sequential algorithm is given by Tibshirani [11] to solve the optimization problem.

While Tibshirani [11] considered estimating coefficients in regression models using LASSO, our interest is in using gene expression data to classify tumors. In particular, we seek to extend the LDA approach advocated by Dudoit et al [7] to handle the case where p is larger than n. We outline the proposed method in the next section.

Estimation methods

We propose to use an optimal scoring procedure for classification, where LASSO estimation is incorporated. In the notation of the previous section, we wish to solve the following optimization problem. Minimize

$$n^{-1}\sum_{i=1}^{n}\{\theta(g_i) - \mathbf{X}_i^T\boldsymbol{\eta}\}^2 + \lambda\sum_{j=1}^{p}|\eta_j| \tag{3}$$

subject to the constraint $N^{-1}\|\mathbf{Z}\boldsymbol{\Theta}\|^2 = 1$. Here is the outline for our procedure.

(1) Choose an initial score matrix $\boldsymbol{\Theta}_0$ satisfying $\boldsymbol{\Theta}_0^T\mathbf{D}_p\boldsymbol{\Theta}_0 = \mathbf{I}$, and let $\boldsymbol{\Theta}_0 = \mathbf{Z}\boldsymbol{\Theta}$.

TABLE 1. Classification error rates (x 100) from simulation study. Numbers in parentheses represent standard errors associated with misclassification rates.

Sample size	$\pi = 0.05$ small effects	$\pi = 0.05$ large effects	$\pi = 0.5$ small effects	$\pi = 0.5$ large effects
$(n_0, n_1) = (15, 15)$	17.3 (1.65)	15.8 (1.63)	12.3 (1.21)	11.9 (1.30)
$(n_0, n_1) = (20, 10)$	20.7 (1.51)	19.3 (1.45)	13.3 (1.35)	12.7 (1.38)
$(n_0, n_1) = (50, 50)$	14.2 (1.15)	13.9 (1.24)	9.8 (1.02)	8.6 (1.11)
$(n_0, n_1) = (70, 30)$	18.3 (1.17)	17.6 (1.29)	10.2 (1.08)	9.9 (1.06)

(2) Fit a linear regression model of $\mathbf{\Theta}_0$ on $\mathbf{X}$ subject to an L_1 constraint on the parameters. Define the fitted values $\mathbf{\Theta}_0^*$. Let $\widehat{\mathbf{f}}(\mathbf{X})$ be the vector of fitted regression functions.

(3) Obtain the eigenvector matrix $\mathbf{\Phi}$ of $\mathbf{\Theta}_0^{*T}\mathbf{\Theta}_0$; the optimal scores are $\mathbf{\Theta} = \mathbf{\Theta}_0\mathbf{\Phi}$.

(4) Define $\mathbf{f}_{\text{opt}}(\mathbf{x}) = \mathbf{\Phi}^T\widehat{\mathbf{f}}(\mathbf{x})$.

Note that we are incorporating the LASSO estimation procedure in step (2) of the algorithm. We cannot use the algorithm of Tibshirani [11] because it is too computationally intensive for large p (number of genes). However, it turns out that the algorithm can be fit using standard software for SVMs, which we will now describe.

Support vector machines

An excellent descriptions of SVMs for classification can be found in [18]. We provide an overview of the method here. We assume that the data are $\{\mathbf{x}_i, y_i\}$ $(i = 1, \ldots, n)$, where $\mathbf{x}_i$ is a d-dimensional vector and $y_i \in \{-1, +1\}$ is the class label. The goal of SVMs is to find an optimal separating hyperplane between the observations with $y = -1$ and those with $y = 1$. This problem can be expressed as minimizing $\|\mathbf{w}\|^2$ subject to the following constraints:

$$\begin{aligned} \mathbf{x}_i \cdot \mathbf{w} + b &\geq 1 - \xi_i \quad \text{for } y_i = 1, \\ \mathbf{x}_i \cdot \mathbf{w} + b &\leq 1 - \xi_i \quad \text{for } y_i = -1, \\ \xi_i &\geq 0 \quad \text{for } i = 1, \ldots, n. \end{aligned} \tag{4}$$

Details on how to solve the optimization problem can be found in [18, chapter 7]. In the unregularized case, fitting the LASSO model is equivalent to fitting an SVM classifier with the following $2p \times 1$ n-dimensional vectors as the inputs: $\mathbf{g}$, $\mathbf{Y}_k$ and $-\mathbf{Y}_k$ $(k = 1, \ldots, p)$, defined to be the sample labels, gene expression values and their negative values for the kth gene across the n samples. The label is the vector $\mathbf{y}_0$, defined to be -1 for the first entry and 1 for the other entries. The proof of the equivalence is given in the "appendix." We have created a macro in R (R foundation) that implements the proposed method and can be obtained from the first author.

As mentioned earlier, an advantage of this approach is that most of the gene effects are estimated to be exactly zero. The method can also identify the genes associated with each of the two classes. Genes whose coefficients are negative are associated with the class $g = -1$, while those with positive estimated coefficients are associated with $g = 1$.

As is evident in the algorithm from the previous section or in (3), the parameter λ needs to be estimated. We use fivefold cross-validation for this.

RESULTS AND DISCUSSION

Simulated data

We first performed a set of simulations to determine how well the proposed methods were at classification. We generated $p = 1000$ dimensional vectors for two populations. We considered the following sample size combinations $(n_0, n_1) = (15, 15)$, $(20, 10)$, $(50, 50)$, and $(70, 30)$, where n_k is the number of samples in the group with $g = k$ $(k = 0, 1)$. All the genes were assumed to be independent with a normal distribution and variance 1. We assumed a model in which a fraction π of the genes was differentially expressed between the two classes, $\pi = 0.05$ and $\pi = 0.5$ were considered. We examined two scenarios. For the first scenario, there was a big change in differential expression in the differentially expressed genes, a shift of 5 units in the mean. In the second scenario, the fold change was only a 1.5 unit difference in mean. For each simulation setting, 100 datasets were generated, and the classification error rates were estimated using three-fold cross-validation. No optimization was performed; we set $\lambda = 10$. The results are summarized in Table 1. Based on the table, we find that for larger sample sizes and larger effect sizes, as well as larger numbers of effects, the error rates are smaller.

However, in our simulations (data not shown), we found that the method had difficulty in selecting the correct variables when p is larger than n. This attests to the fact that variable selection in the situation of large p and small n is quite difficult. We discuss this situation in the "conclusion."

Table 2. List of genes underexpressed in prostate cancer relative to benign prostate tissue.

Clone ID	Gene name
Hs.288965	*Homo sapiens* cDNA: FLJ22300 fis, clone HRC04759
Hs.76307	Neuroblastoma, suppression of tumorigenicity 1
Hs.9615	Myosin, light polypeptide 9, regulatory
Hs.226795	Glutathione S-transferase pi
Hs.171731	Solute carrier family 14 (urea transporter), member 1 (Kidd blood group)

Prostate cancer gene expression data

The example we consider is from a prostate cancer study; a subset of the samples was considered by Dhanasekaran et al [3]. We focus here on noncancer versus cancer tissues. The samples are profiled using spotted cDNA (ie, red/green) microarrays; there are initially 101 samples profiled using 10 K chips (9984 genes). We have taken the following preprocessing steps:

(1) remove genes that are reported as missing in more than 10% of the samples;

(2) remove genes that have a variance less than 0.05 in all samples;

(3) impute measurements for missing genes using the median.

This leaves a total of 4880 genes for analysis.

We first performed an estimation of the error rate using fivefold cross-validation. This generally gave an error rate between 15–20% for various choices of λ, suggesting that the classifier is not sensitive to the choice of the smoothing parameter.

One of the by-products of the procedure is a list of genes that are estimated to have non-zero effects. We present the gene lists for $\lambda = 1$ in Table 2. Out of the 4880 genes, only 21 are estimated to have nonzero effects. Of the genes that are overexpressed in prostate cancer relative to benign prostate tissue, the early growth response (Hs. 326035/301865), feline sarcoma viral oncogene homolog (Hs.81665), T-cell receptor gamma locus (Hs. 112259), and fatty acid synthase (Hs.83190) have been seen by other investigators to be upregulated in prostate cancer, as in Table 3. The other genes on the list could represent false positives or genes whose joint effect is predictive of cancer status.

Conclusion

In this paper, we have introduced a new approach to the joint problems of classification and variable selection in the analysis of microarray data. These problems have been treated as separate problems in the previous literature. Our approach is combine the two problems by use of the LASSO.

This work has opened the way for several future avenues of research that we are currently investigating. First, a popular alternative to LDA in classification problems is logistic regression. It has been recently motivated by ROC considerations (McIntosh and Pepe [19]). While it is possible to formulate a LASSO estimation for logistic regression models, adapting the LASSO-SVM equivalence to this situation requires new algorithms. It will also be important to compare the performance of the two L_1-regularized procedures (LDA and logistic regression) on real and simulated microarray datasets.

In this paper, we focused on the two-class problem. While LDA and logistic regression can be extended to accommodate multicategorical responses, the ROC arguments that motivated the method here only exist for two populations. We are currently exploring theoretical frameworks for generalizing ROC ideas for multiple disease states.

The estimation procedure described in this paper allows the joint estimation of multivariate gene effects on the response (class label). The approach described here could be generalized by fitting more nonlinear gene effects in the estimation algorithm or by including higher-order interactions between genes. Another generalization is to perform a clustering of the genes and to enter the cluster averages as covariates in the model. Such an approach was taken by Hastie et al [20] and Tibshirani et al [21].

It is also of current interest to incorporate biological knowledge into microarray data analyses. In many instances, scientists are interested in the effects of a particular gene or pathway on genetic expression. In this context, approaches have been suggested by Zien et al [22] and Pavlidis et al [23] in which biological knowledge as represented by pathway scores or functional annotation status are correlated with gene expression. However, their approaches were univariate. There would be potential gains in efficiencies of analyses by considering joint models for pathways. We are currently studying the applicability of the joint estimation procedure described here to that setting.

Finally, a by-product of the method proposed here is that the individual genes can be estimated to have exactly zero effect on the response. The list of genes with estimated nonzero effects then comprise a gene list that

Table 3. List of genes overexpressed in prostate cancer relative to benign prostate tissue.

Clone ID	Gene name
Hs.326035/301865	Early growth response 1 -OR- dopachrome tautomerase (dopachrome delta-isomerase, tyrosine-related protein 2)
Hs.299221	Pyruvate dehydrogenase kinase, isoenzyme 4
Hs.81665	v-kit Hardy-Zuckerman 4 feline sarcoma viral oncogene homolog
Hs.74267	Ribosomal protein L15
Hs.75431	Fibrinogen, gamma polypeptide
Hs.335797	ESTs, moderately similar to hypothetical protein FLJ20097 (*Homo sapiens*) (*H sapiens*)
Hs.82129	Carbonic anhydrase III, muscle specific
Hs.112259	T-cell receptor gamma locus
Hs.151258	Hypothetical protein FLJ21062
Hs.22394	Sec3-like
Hs.84190	Solute carrier family 19 (folate transporter), member 1
Hs.119597	Stearoyl-CoA desaturase (delta-9-desaturase)
Hs.131740	*Homo sapiens* cDNA FLJ30428 fis, clone BRACE2008941
Hs.50727	N-acetylglucosaminidase, alpha- (Sanfilippo disease IIIB)
Hs.83190	Fatty acid synthase
Hs.82961	*Homo sapiens*, clone MGC: 22588 IMAGE: 4696566, mRNA, complete cds

investigators can do further validation work on. However, in our simulations (data not shown), we found that the method had difficulty in selecting the correct v variables. This attests to the fact that variable selection in the situation of large p and small n is quite difficult. An alternative to the method proposed here is Bayesian variable selection methods (Lee et al [24]). We are currently exploring an adaptation of the algorithm described here to a Bayesian approach.

APPENDIX

If we let $\mathbf{w} = (w_1, \ldots, w_p)$, then SVMs can be shown to minimize $\|\mathbf{w}\|^2$ among all hyperplanes with norm 1, subject to the constraint that $g_i(\mathbf{w} \cdot \mathbf{x}_i + b) \geq 1$ for all $i = 1, \ldots, n$. The quantity $2/\|\mathbf{w}\|$ is known as the margin. In other words, we are trying to find the separating hyperplane that maximizes the margin among all classifiers that satisfy the inequality constraints. Using Lagrange multipliers, we can formulate the optimization problem as finding $\mathbf{w}$ and b to minimize

$$L(\mathbf{w}, b) \equiv \frac{1}{2}\|\mathbf{w}\|^2 - \sum_{i=1}^{n} \gamma_i g_i\big(\langle \mathbf{x}_i, \mathbf{w}\rangle + b\big) + \boldsymbol{\gamma}'\mathbf{1}, \qquad \text{(A.1)}$$

subject to $\gamma_i \geq 0$ $(i = 1, \ldots, n)$, where $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_n)$. Instead, we consider the dual of this problem, which is to maximize L such that the derivatives with respect to $\mathbf{w}$ and b vanish and also that $\gamma_i \geq 0$ $(i = 1, \ldots, n)$. By differentiating (A.1) with respect to $\mathbf{w}$ and b and setting the resulting derivatives equal to $\mathbf{0}$, we obtain

$$\begin{aligned} \frac{\partial L}{\partial \mathbf{w}} &= \mathbf{w} - \sum_{i=1}^{n} \gamma_i g_i \mathbf{x}_i = \mathbf{0}, \\ \frac{\partial L}{\partial b} &= -\sum_{i=1}^{n} \gamma_i g_i = 0. \end{aligned} \qquad \text{(A.2)}$$

Equations (A.2) yield the solutions $\hat{\mathbf{w}} = \sum_{i=1}^{n} \gamma_i g_i \mathbf{x}_i$ and $\sum_{i=1}^{n} \gamma_i g_i = 0$. If we plug in the formula for $\hat{\mathbf{w}}$ into (A.1), the optimization problem becomes one of maximizing the dual function $W(\boldsymbol{\eta})$ over $\boldsymbol{\gamma} \geq \mathbf{0}$ and $\sum_{i=1}^{n} \gamma_i g_i = 0$, where

$$W(\boldsymbol{\eta}) = \sum_{j=1}^{n} \gamma_j - \frac{1}{2} \sum_{j,k=1}^{n} \gamma_j \gamma_k g_j g_k \langle \mathbf{x}_j, \mathbf{x}_k \rangle. \qquad \text{(A.3)}$$

Tibshirani [11] considered the following estimation problem Minimize

$$\sum_{i=1}^{n} \left(Y_i - \mathbf{Z}_i^T \beta\right)^2 \qquad \text{(A.4)}$$

subject to $\sum_{j=1}^{p} |\beta_j| \leq t$. Note that this minimization problem is equivalent to minimizing (A.4) subject to $\sum_{j=1}^{p} (\beta_j^+ + \beta_j^-) \leq t$, where $a^+ = \max(0, a)$ and $a^- = -\min(0, -a)$. We can equivalently consider minimization of

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=1}^{p} Z_{ij} \beta_j^+ + \sum_{j=1}^{p} Z_{ij} \beta_j^- \right)^2 - C\left[t - \sum_{j=1}^{p} \beta_j^+ - \sum_{j=1}^{p} \beta_j^- \right] \qquad \text{(A.5)}$$

subject to $\beta_j^+ \geq 0$ and $\beta_j^- \geq 0$, $j = 1,\dots,p$. We introduce some more notation. For $k = 1,\dots,2p$, define W_{ik} as Z_{ik} for $k = 1,\dots,p$ and $-Z_{i(k-p-1)}$ for $k = p+1,\dots,2p$. Similarly, define the $2p \times 1$ dimensional vector $\boldsymbol{\eta} = (\eta_1,\eta_2,\dots,\eta_{2p})$ by $\eta_j = \beta_j^+$ for $j = 1,\dots,p$ and $\eta_j = \beta_{j-p-1}^-$ for $j = p+1,\dots,2p$. Thus, (A.5) can be written as

$$\sum_{i=1}^{n}\left(Y_i - \sum_{j=1}^{2p} W_{ij}\eta_j\right)^2 - C\left[t - \sum_{j=1}^{2p}\eta_j\right]. \tag{A.6}$$

The optimization problem now is to minimize (A.6) subject to $\eta_j \geq 0$ for $j = 1,\dots,2p$. Expanding the squared term in (A.6), we have

$$\begin{aligned}\sum_{i=1}^{n}&\left(Y_i^2 - 2Y_i\sum_{j=1}^{2p} W_{ij}\eta_j - \sum_{j,k=1}^{2p}\eta_j\eta_k W_{ij}W_{ik}\right)\\&\qquad - C\left[t - \sum_{j=1}^{2p}\eta_j\right].\end{aligned} \tag{A.7}$$

Distributing the summation sign and interchanging indices, (A.7) is equivalent to

$$\begin{aligned}&\langle Y, Y\rangle - 2\sum_{j=1}^{2p}\langle W_j, Y\rangle\eta_j\\&\qquad + \sum_{j,k=1}^{2p}\eta_j\eta_k\langle W_j, W_k\rangle - C\left[t - \sum_{j=1}^{2p}\eta_j\right].\end{aligned} \tag{A.8}$$

In particular, we want to minimize (A.8).

We now reconsider the optimization problem (A.3). Suppose we define new observations $(g_i, \mathbf{x}_i)$ $(i = 1,\dots,2p+1)$ by $g_1 = -1$ and $g_j = 1$ for $j = 2,\dots,2p+1$, $\mathbf{x}_1 = Y/t$, and $\mathbf{x}_j = W_{j-1}$ for $j = 2,\dots,2p+1$ and parameters $(\gamma_1,\dots,\gamma_{2p+1})$ by

$$\gamma_1 = \frac{2t^2}{\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{2p} W_{ij}\eta_j\right)^2} \tag{A.9}$$

and $\gamma_j = \alpha_1\eta_{j-1}/t$ for $j = 2,\dots,2p+1$. Then the condition $\sum_{i=1}^{2p+1}\gamma_i g_i = 0$ is equivalent to $\gamma_1 = \sum_{i=2}^{2p+1}\gamma_i$, which after further algebraic simplification, yields $\sum_{j=1}^{2p}\eta_j = t$. Considerable algebraic simplification gives that maximizing (A.3) can be rewritten as a problem of maximizing

$$\begin{aligned}&2\alpha_1 - \frac{1}{2}\frac{\alpha_1^2}{t^2}\langle Y, Y\rangle + \frac{\alpha_1^2}{t^2}\sum_{j=1}^{2p}\eta_j\langle W_j, Y\rangle\\&\qquad - \frac{1}{2}\frac{\alpha_1^2}{t^2}\sum_{j,k=1}^{2p}\eta_j\eta_k g_j\langle W_j, W_k\rangle\end{aligned} \tag{A.10}$$

subject to $\boldsymbol{\eta} \geq 0$ and $\sum_{j=1}^{2p}\eta_j = t$. Because $\alpha_1 \geq 0$, comparison of problems (A.10) and (A.8) reveal that they should yield the same solution.

ACKNOWLEDGMENT

The research of the first author was supported by grant NIH 1R01GM72007-01 from the Joint DMS/DBS/NIGMS Biological Mathematics Program.

REFERENCES

[1] Alizadeh AA, Ross DT, Perou CM, van de Rijn M. Towards a novel classification of human malignancies based on gene expression patterns. *J Pathol.* 2001;195(1):41–52.

[2] Khan J, Wei JS, Ringner M, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nat Med.* 2001;7(6):673–679.

[3] Dhanasekaran SM, Barrette TR, Ghosh D, et al. Delineation of prognostic biomarkers in prostate cancer. *Nature.* 2001;412(6849):822–826.

[4] Efron B, Tibshirani R, Storey JD, Tusher V. Empirical Bayes analysis of a microarray experiment. *J Am Stat Assoc.* 2001;96(456):1151–1160.

[5] Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA.* 1998;95(25):14863–14868.

[6] West M. Bayesian factor regression models in the "large p, small n" paradigm. In: *Bayesian Statistics 7 Proceedings of the Seventh Valencia International Meeting.* New York, NY: Oxford University Press; 2003:723–732.

[7] Dudoit S, Fridlyand J, Speed TP. Comparison of discrimination methods for the classification of tumors using gene expression data. *J Am Stat Assoc.* 2002;97(457):77–87.

[8] van't Veer LJ, Dai H, van de Vijver MJ, et al. Gene expression profiling predicts clinical outcome of breast cancer. *Nature.* 2002;415(6871):530–536.

[9] van de Vijver MJ, He YD, van't Veer LJ, et al. A gene-expression signature as a predictor of survival in breast cancer. *N Engl J Med.* 2002;347(25):1999–2009.

[10] Nguyen DV, Rocke DM. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics.* 2002;18(1):39–50.

[11] Tibshirani RJ. Regression shrinkage and selection via the LASSO. *J Roy Statist Soc B.* 1996;58(1):267–288.

[12] Su JQ, Liu JS. Linear combinations of multiple diagnostic markers. *J Am Stat Assoc.* 1993;88:1350–1355.

[13] Baker SG. Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics.* 2000;56(4):1082–1087.

[14] Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics.* 2000;1(2):123–140.

[15] Bamber D. The area above the ordinal dominance graph and the area below the receiver operating

characteristic graph. *J Math Psych.* 1975;12(4):387–415.

[16] Hastie T, Tibshirani R, Buja A. Flexible discriminant analysis by optimal scoring. *J Am Stat Assoc.* 1994;89(428):1255–1270.

[17] Hastie T, Buja A, Tibshirani R. Penalized discriminant analysis. *Ann Statist.* 1995;23(1):73–102.

[18] Cristianini N, Shawe-Taylor J. *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods.* Cambridge: Cambridge University Press; 2000.

[19] McIntosh MW, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics.* 2002;58(3):657–664.

[20] Hastie T, Tibshirani R, Eisen MB, et al. "Gene shaving" as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol.* 2000;1(2):Research0003. Epub 2000 Aug 04.

[21] Tibshirani R, Hastie T, Narasimhan B, et al. Exploratory screening of genes and clusters from microarray experiments. *Statist Sinica.* 2002;12(1):47–59.

[22] Zien A, Kuffner R, Zimmer R, Lengauer T. Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol.* 2000;8:407–417.

[23] Pavlidis P, Lewis DP, Noble WS. Exploring gene expression data with class scores. *Pac Symp Biocomput.* 2002;7:474–485.

[24] Lee KE, Sha N, Dougherty ER, Vannucci M, Mallick BK. Gene selection: a Bayesian variable selection approach. *Bioinformatics.* 2003;19(1):90–97.