Biostatistics (2003), **4**, 3, *pp.* 449–463
Printed in Great Britain

A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection

YUTAKA YASUI[†]

Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109-1024, USA
yyasui@fhcrc.org

MARGARET PEPE, MARY LOU THOMPSON

Department of Biostatistics, University of Washington, Seattle, WA, USA

BAO-LING ADAM, GEORGE L. WRIGHT, JR.

Department of Microbiology and Molecular Cell Biology and Virginia Prostate Center, Eastern Virginia Medical School, Norfolk, VA, USA

YINSHENG QU, JOHN D. POTTER, MARCY WINGET, MARK THORNQUIST, ZIDING FENG

Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., Seattle, WA 98109-1024, USA

SUMMARY

With recent advances in mass spectrometry techniques, it is now possible to investigate proteins over a wide range of molecular weights in small biological specimens. This advance has generated data-analytic challenges in proteomics, similar to those created by microarray technologies in genetics, namely, discovery of 'signature' protein profiles specific to each pathologic state (e.g. normal vs. cancer) or differential profiles between experimental conditions (e.g. treated by a drug of interest vs. untreated) from high-dimensional data. We propose a data-analytic strategy for discovering protein biomarkers based on such high-dimensional mass spectrometry data. A real biomarker-discovery project on prostate cancer is taken as a concrete example throughout the paper: the project aims to identify proteins in serum that distinguish cancer, benign hyperplasia, and normal states of prostate using the Surface Enhanced Laser Desorption/Ionization (SELDI) technology, a recently developed mass spectrometry technique.

Our data-analytic strategy takes properties of the SELDI mass spectrometer into account: the SELDI output of a specimen contains about 48 000 (x, y) points where x is the protein mass divided by the number of charges introduced by ionization and y is the protein intensity of the corresponding mass per charge value, x, in that specimen. Given high coefficients of variation and other characteristics of protein intensity measures (y values), we reduce the measures of protein intensities to a set of binary variables that indicate peaks in the y-axis direction in the nearest neighborhoods of each mass per charge point in the x-axis direction. We then account for a shifting (measurement error) problem of the x-axis in SELDI output. After this pre-analysis processing of data, we combine the binary predictors to

[†]To whom correspondence should be addressed

generate classification rules for cancer, benign hyperplasia, and normal states of prostate. Our approach is to apply the boosting algorithm to select binary predictors and construct a summary classifier. We empirically evaluate sensitivity and specificity of the resulting summary classifiers with a test dataset that is independent from the training dataset used to construct the summary classifiers. The proposed method performed nearly perfectly in distinguishing cancer and benign hyperplasia from normal. In the classification of cancer vs. benign hyperplasia, however, an appreciable proportion of the benign specimens were classified incorrectly as cancer. We discuss practical issues associated with our proposed approach to the analysis of SELDI output and its application in cancer biomarker discovery.

Keywords: Bioinformatics; Classification; Disease markers; Machine Learning; Mass spectrometry.

1. INTRODUCTION

The central dogma of molecular biology states that proteins are closer to actual biologic functions of cells than mRNAs or DNAs (Alberts *et al.*, 1994). This argues for seeking protein biomarkers of a disease, in addition to genetic biomarkers. With recent advances in mass spectrometry techniques, it is now possible to investigate proteins over a wide range of molecular weights in small biological specimens, e.g. serum (Rubin and Merchant, 2000; Srinivas *et al.*, 2001). This advance has generated data-analytic challenges in proteomics, similar to those created by microarray technologies in genetics (Lander, 1999; Liotta and Petricoin, 2000), namely discovery of 'signature' profiles specific to each pathologic state (e.g. malignant, benign, or normal) or differential profiles between experimental conditions (e.g. treated or untreated by a drug of interest) from high-dimensional data. This paper proposes an analytic strategy for discovering protein biomarker profiles based on such high-dimensional mass spectrometry data.

2. THE PROSTATE CANCER PROTEIN BIOMARKER DISCOVERY PROJECT

Background

Our data-analytic strategy was motivated by a biomarker discovery project on prostate cancer carried out at the Department of Microbiology and Molecular Cell Biology and Virginia Prostate Center of the Eastern Virginia Medical School (EVMS). The project is part of a large National-Cancer-Institute-funded research consortium, the Early Detection Research Network (Srivastava and Kramer, 2000). The overall goal of the Network is to discover, and validate clinically, new biomarkers of cancer that enable earlier detection and, consequently, better survival and cure rates. The ultimate goal of the EVMS research project is to identify serum protein biomarkers of prostate cancer, benign prostatic hyperplasia (BPH), and normal, and distinguish them from each other. The basis for the protein-based early detection of cancer is the concept that a transformed cancerous cell and its clonal expansion would result in up- (or down-) regulation of certain proteins: our aim is to identify such early molecular signs of prostate cancer by measuring protein profiles in serum. The goal of the analysis is to assess whether protein profiles can distinguish the three disease groups as defined, and identify signature profiles for the classification. This allows researchers to identify proteins/peptides associated with the signature profiles and study their biological significance, which eventually leads to a clinical detection tool.

Serum samples and their protein analyses

The discussion in this paper will focus on the first stage of the protein biomarker discovery/validation process, namely, an exploratory data-driven identification of protein profiles that appear to distinguish prostate cancer cases from cancer-free subjects (those with BPH and normals); see Pepe *et al.* (2001) for

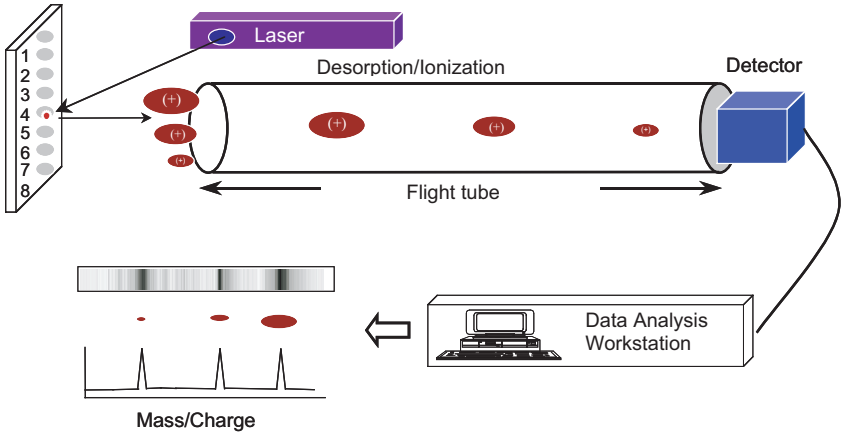

Fig. 1. The outline of the Surface-Enhanced Laser Desorption/Ionization (SELDI) ProteinChip technology.

subsequent stages of the biomarker discovery/validation process. For this purpose, we retrieved serum samples of 386 subjects, approximately equally distributed across four groups: late-stage prostate cancer ($N = 98$), early-stage prostate cancer ($N = 99$), BPH ($N = 93$), and normal controls ($N = 96$), stored in the serum repository of the EVMS Virginia Prostate Center. Late-stage prostate cancer patients had a biopsy-proven cancer staged C or D, and Prostate Specific Antigen (PSA) concentrations greater than 4 ng/ml. Early-stage prostate cancer patients had a biopsy-proven cancer staged A or B, and PSA greater than 4 ng/ml. BPH patients had PSA values between 4 ng/ml and 10 ng/ml, low PSA velocities, and at least two negative biopsies. Normal controls were aged 50 or older, corresponding to the age range of cancer and BPH patients, with a PSA level less than 4 ng/ml and normal digital rectal exam.

Each of the retrieved serum samples was assayed at the EVMS for protein expression by the SELDI ProteinChip Array technology of Ciphergen Biosystems Inc. (Wright *et al.*, 1999; Rubin and Merchant, 2000; Adam *et al.*, 2001; Srinivas *et al.*, 2001). The SELDI technology is a time-of-flight mass spectrometry, shown in Figure 1, with a special ProteinChip Array whose surface captures proteins using chemically or biologically defined protein docking sites. Proteins are captured on the chip surface, purified by washing the surface, and crystallized with small molecules called 'matrix' or 'energy-absorbing molecules' (EAMs) whose function is to absorb laser energy and transfer it to proteins. Energized protein molecules fly away from the surface into a time-of-flight tube where the time for the molecules to fly through the tube is a function of the molecular weight and charge of the protein. A detector at the end of the tube measures the 'intensity' of proteins at each discrete time of flight and outputs about 48 000 data points of (time of flight, intensity) pairs. Each discrete time of flight corresponds uniquely to a ratio of the molecular weight of a protein to the number of charges introduced by the ionization. SELDI output, therefore, produces about 48 000 data points of (mass/charge, intensity) pairs. Our analyses used 11 175 data points per sample covering the mass/charge range of 1500–20 000. The lower limit of the range was placed at 1500 because intensity measures of proteins below this limit are distorted by those of EAMs. Twenty thousand was considered to be a reasonable upper limit of the range where the SELDI assays are most sensitive. Note that PSA, a 28 741-Da glycoprotein that is currently used in the screening of prostate cancer, is out of the mass/charge range of 1500–20 000 unless the molecules were doubly charged (an uncommon event).

The design of the protein-marker discovery analysis

The SELDI output from the 386 serum samples was separated by a stratified random sampling into 'test data' (a total of 60, 15 samples from each of the four groups) and 'training data' (a total of 326 samples). Our aim was to develop a data-analytic strategy using the training data, for which both the true pathologic state and SELDI output of the 11 175 (mass/charge, intensity) data pairs of each sample were available. The resulting data-analytic strategy was then applied to the test data for estimating the true classification proportion under each pathologic state. In the testing samples, the true pathologic state was blinded except to a statistician who was not involved in the development of the data-analytic strategy and it was this statistician who provided estimated true classification proportions.

3. PRE-ANALYSIS PROCESSING OF SELDI OUTPUT

Our quality-control experiments suggested several measurement properties of SELDI output which must be accounted for in the analysis. (Note that these properties may change by experimental condition, and strategies for standardizing these properties are currently under active development.) First, the coefficient of variation (CV) of intensity measures is approximately 50–60%. Thus, there are substantial measurement errors in the absolute values of intensity at any given mass/charge points. Second, in a simpler experiment in which the samples are controlled to contain only a few proteins, the CV was approximately halved if 'relative' values of intensity (i.e. intensity measures divided by an intensity measure at a given mass) rather than 'absolute' values of intensity were considered. Third, an intensity value is a function of the laser energy applied and mass/charge for which the intensity was measured, given amounts of proteins being constant. The intensity value at any given mass/charge point increases if higher laser energy is applied. With a certain laser energy applied, the intensity is generally higher for lower mass/charge points. Fourth, the mass-axis of the SELDI output shifts from experiment to experiment by approximately $\pm$ 0.1–0.2% of the mass/charge value. Consideration of these measurement properties of our experimental condition led us to the following strategy for data analysis.

Reduction of intensity measures into binary signals

Because of the high CV of the original intensity measures of SELDI output, we chose not to rely on the absolute intensity values themselves for establishing biomarkers. Instead, we decided to reduce the absolute intensity measures into local peak/non-peak binary data. This was accomplished by first assessing, at each mass/charge point, whether or not the intensity at that point is the highest among its nearest $\pm N$-point neighborhood set, nearest with respect to the mass/charge axis: $N = 10, 20, 30, 40$ were initially considered and $N = 20$ was chosen. Figure 2 shows an example of SELDI output in the mass/charge range of 9000–11 000: the highest-intensity points among its nearest $\pm N$-point neighborhood set were marked by '$\times$' in the four plots using $N = 10$, 20, 30, and 40, respectively. This example shows that the processing with $N = 30$ or 40 misses some visually apparent peaks (e.g. those around mass/charge of 9500), while the processing with $N = 10$ identifies many peaks that may be random noise (e.g. those around mass/charge of 10 000). We considered $N = 20$ as the best choice after visually examining multiple plots similar to Figure 2.

Although this initial step worked fairly well in identifying visually apparent peaks, some points that seemed clearly random noise were also identified as peaks (those marked by '$\times$' but with near-zero intensity values that are not clearly distinguishable from neighborhood random fluctuation of intensity). We, therefore, considered an additional criterion for the definition of a peak. Specifically, a peak must have an intensity value that is higher than an 'average' intensity level of its *broad* neighborhood. The average intensity of the broad neighborhood was calculated by the super-smoother method (Friedman,

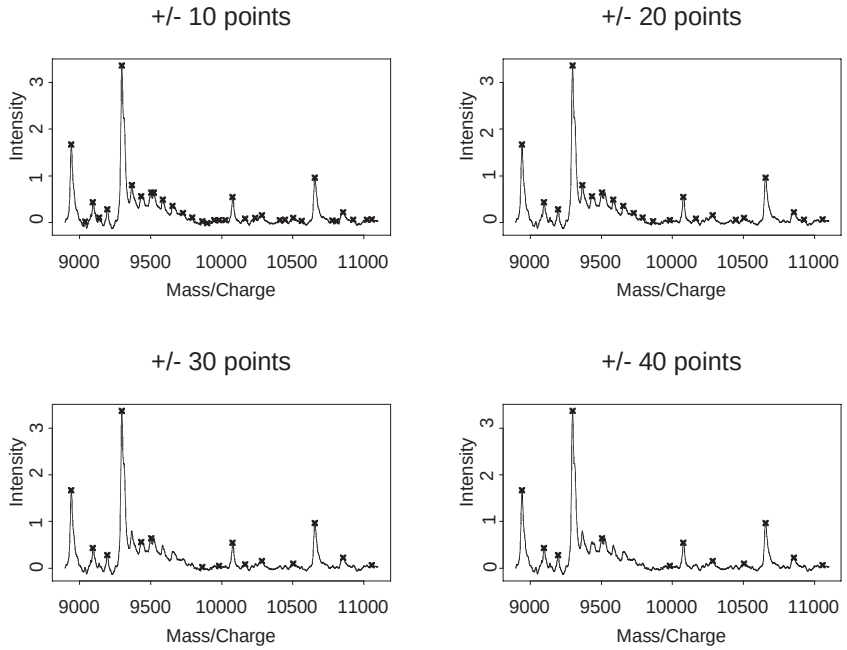

Fig. 2. Pre-analysis processing of SELDI output for a sample in the mass/charge range of 9000–11 000. Each mass/charge point with an intensity that is highest among its nearest N-point neighborhood set is marked by '$\times$': $N = \pm 10$, 20, 30, 40 points were considered.

1984) using 5% of all data points as the smoothing window. Figure 3 shows 'peaks' marked by '$\otimes$', the points that have intensity values that are higher than the averages in their broad neighborhoods and also the highest in their respective nearest $\pm$ 20-point neighborhood sets. Similar to the selection of $N = 20$, we selected the window width as 5% of all data points in the super-smoother, empirically by trial and error with visual checking of the resulting peak/non-peak data with the original plots of the intensity.

The above processing is consistent with the measurement properties of SELDI output: (1) we reduced the continuous intensity measures into less quantitative binary data because of their high CVs; and (2) the peak/non-peak at each mass/charge was determined locally since the intensity is a function of mass/charge given the laser energy level and the amount of protein. Note that the peak/non-peak at each mass/charge should not be interpreted as a presence/absence indicator of the protein corresponding to that mass: if any of the nearest neighborhood points or the smoothed average over the 5% data window is higher than the intensity of the point in question, then the point is classified as a non-peak even if the protein is present and its intensity is non-zero.

Mass/charge alignment

To alleviate the impact of the mass/charge axis shifting problem (i.e. the error of approximately $\pm$0.1–0.2% of the mass/charge value), we chose to label all the points within $\pm$0.2% of the mass/charge value of each peak point as peaks (see horizontal bars of peaks in Figure 3). That is, after identifying each peak as a point by the above procedure, we created an interval of mass/charge values, whose points are all marked as peaks, with the width of the interval being the 0.4% of the mass/charge value of its mid-point, the originally identified peak point. Undoubtedly, many of the mass/charge points in such a peak

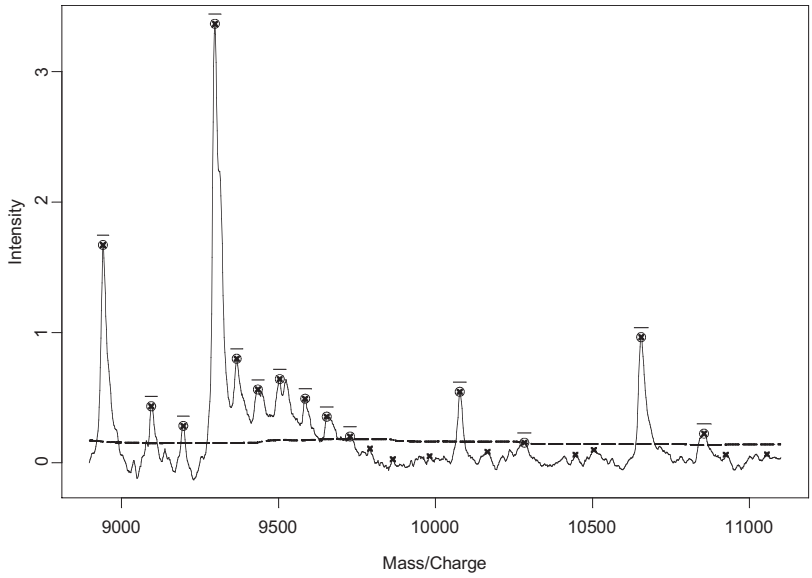

Fig. 3. An example of SELDI data after the pre-analysis processing. The broken line indicates the average intensity calculated by the super-smoother method using 5% of all data points (i.e. 5% of 11 175 points) as the smoothing window: the figure shows the region corresponding to about 10% of the 11 175 points. The points marked by '⊗' are the peaks, points that have intensity values that are higher than the averages in their broad neighborhoods and also the highest in their respective nearest ±20-point neighborhood sets. A horizontal bar above each peak shows an interval corresponding ±0.2% of the peak's mass/charge value.

interval are falsely labeled as peaks. The aim of this crude approach was, however, to keep the number of false-negative points that are created by the mass/charge axis shifting problem small. If the resulting peak intervals are sufficiently distant in the mass/charge axis direction and mutually exclusive around real biomarker peaks (i.e. peaks that are specific to a certain pathologic state), the false-positive peaks will not interfere with the search for real biomarker peaks when comparing peak profiles across pathologic states. In other words, the proposed approach has a limitation in that it will not perform well for biomarker peaks that are close to any non-biomarker peak in the mass/charge axis direction.

An alternative approach to the mass/charge axis shifting problem was tested, but rejected. We tried to identify a mass/charge point (or a set of mass/charge points) that is (are) consistently a peak (peaks) in every SELDI output and align the mass/charge axis based on this (these) 'anchor' point(s). However, we were not able to find any mass/charge point that could serve as a reliable anchor for this purpose.

4. AN APPLICATION OF THE BOOSTING ALGORITHM TO THE CONSTRUCTION OF A CLASSIFIER

Following the processing of SELDI output described above, we constructed classifiers using only a small subset of mass/charge points to separate the pathologic states. To this end, we employed a relatively new classifier-building methodology called 'boosting'.

Boosting algorithm

Boosting is an ingenious method, proposed by Schapire and Freund (Schapire, 1990; Freund, 1995; Freund and Schapire, 1996), for combining 'weak' base classifiers $(c_1, c_2, c_3, \ldots)$ into a 'powerful' summary classifier $(s_M = c_1 + c_2 + c_3 + \cdots + c_M)$. The algorithm performs a weighted stage-wise selection of a base classifier given all the previously selected base classifiers. In selecting the next base classifier, c_m, at the mth stage, higher weights are given to samples that are incorrectly classified by the current summary classifier, s_{m-1}, so that c_m will be selected with a tendency towards correctly classifying the previously incorrectly classified samples. The method has been shown to work well in a wide range of classification problems. See Friedman *et al.* (2000) and Hastie *et al.* (2001) for descriptions and discussion of boosting: specifically, they showed boosting as an approximate maximum-likelihood fitting algorithm for additive logistic models, providing a useful statistical understanding of the algorithm.

We apply boosting to construct a summary classifier, s_M, of the following form that is based on the peak/non-peak information at M mass/charge points:

$$s_M = \sum_{i=1}^{M} c_i = \sum_{i=1}^{M} (\alpha_i + \beta_i X_i)$$

where α_i and β_i are parameters and each X_i is a binary peak indicator ($X_i = 1$ for a peak and $X_i = 0$ otherwise) at a certain mass/charge, m_i. To select which mass/charge should enter the summary classifier and determine the associated parameter estimates α_i and β_i we used a generalized version of popular AdaBoost procedure (Freund and Schapire, 1996), called Real AdaBoost (Friedman *et al.*, 2000). Specifically, our boosting algorithm proceeds as follows. For reasons to be described in Section 5, we consider a classification problem of two groups at a time, rather than simultaneous classification of all pathologic states. Let the two groups be indexed by $Y = 1$ and $Y = 0$, and suppose a total of N samples are available in the training dataset. We start with an equal weight for each sample, $w_1 = 1/N$, and select the most statistically significant mass/charge, m_1, by a likelihood-ratio test for a linear logistic regression model for the binary outcome Y (with weights w_1). The linear predictor of the logistic regression model forms the 'best' base classifier $c_1 = \alpha_1 + \beta_1 X_1$ ($c_1 < 0$ predicts $Y = 0$ and $c_1 \geqslant 0$ predicts $Y = 1$) in this first stage. Note that α_1 and $(\alpha_1 + \beta_1)$ represent the log odds of $Y = 1$ for samples with $X_1 = 0$ (non-peak) and $X_1 = 1$ (peak), respectively, at mass/charge m_1. Of all the mass/charge values that could be selected into the model, m_1 is associated with the highest significance by the likelihood-ratio test for testing $\beta_1 = 0$. We then update the weight for each sample by $w_2 = w_1 \times \exp(|0.5 s_1|)$ if $s_1 (= c_1)$ indicates the incorrect classification for the sample or $w_2 = w_1 \times \exp(-|0.5 s_1|)$ if s_1 indicates the correct classification for the sample: see Friedman *et al.* (2000) for an interpretation of this form of weights. Using the new weights w_2 for the logistic regression, we choose the most significant mass/charge, m_2, by a likelihood-ratio test. Note that X_1 selected in the first stage does not enter the logistic regression as an explanatory variable in the second stage: the only influence of X_1 on the second-stage selection is through the weights w_2. The linear predictor of the logistic regression model forms the 'best' base classifier $c_2 = \alpha_2 + \beta_2 X_2$ in the second stage, and constructs a summary classifier, $s_2 = c_1 + c_2$: $s_2 < 0$ predicts $Y = 0$ and $s_2 \geqslant 0$ predicts $Y = 1$. We continue by updating weights $w_3 = w_2 \times \exp(+|0.5 s_2|)$ if s_2 is incorrect or $w_3 = w_2 \times \exp(-|0.5 s_2|)$ if s_2 is correct. This iterative process is repeated without any pre-specified limit. The final classification rule at the Mth stage is: $s_M < 0$ predicts $Y = 0$ and $s_M \geqslant 0$ predicts $Y = 1$.

Note that the boosting summary classifier, s_M, can be written as the linear predictor of a logistic regression model for the binary outcome variable Y with $(M + 1)$ parameters:

$$\text{logit}\{P(Y = 1)\} = s_M = \sum_{i=1}^{M} (\alpha_i + \beta_i X_i) = \left\{ \sum_{i-1}^{M} (\alpha_i) \right\} + \sum_{i=1}^{M} \beta_i X_i .$$

The boosting approach is similar to linear logistic regression with a forward variable selection in that an X_i is selected at each stage and the selection influences its subsequent stages. There are fundamental differences, however. The boosting approach estimates only α_i and β_i at the ith stage keeping all the previous parameter estimates unchanged: the influence of the previous stages on the subsequent stages is carried through by the weights that are determined at each stage according to the classification performance of that stage. In linear logistic regression with a forward variable selection, on the other hand, the parameters are estimated simultaneously without weighting by the maximum likelihood method at each stage and the estimates change from stage to stage: the influence of the previous stages on the subsequent stages is carried through by the set of predictors selected into the model in the previous stages.

A stopping rule for boosting

Boosting has been shown empirically to be highly resistant to overfitting. Many real data examples demonstrate that even a large number of boosting iterations does not lead to a decrease in the classification ability of the summary classifier (Friedman *et al.* 2000). In fact, the performance of a summary classifier constructed by boosting often converges as the iteration number, M, increases. This is in contrast to standard classification methods, such as logistic discriminant analysis, where use of too many predictor variables (i.e. protein mass/charge points in our case) in the classifier lowers its classification performance. This important property of boosting, however, has not been fully explained (Friedman *et al.*, 2000) and a few counter-examples have been given (Friedman *et al.*, 2000; Ridgeway, 2000).

There are two reasons for stopping boosting iterations in our application. First, parsimony of the summary classifier is advantageous in studying basic biological properties of the set of proteins involved in the classifier. Additional boosting iterations that do not lead to a significant improvement in the performance of the summary classifier are not warranted. Second, although the empirical evidence supports the idea that boosting is highly resistant to overfitting, there are counter-examples and no theoretical guidance exists as to when overfitting may occur.

We considered a simple stopping rule for boosting iterations here: specifically, we stop the boosting iterations when both observed sensitivity and specificity in the training dataset exceed certain minimum values (e.g. $>90\%$ sensitivity and $>95\%$ specificity). We then take the boosting summary classifier at the stopped iteration and apply it to the test dataset for the empirical assessment of sensitivity and specificity. Alternatively, the boosting algorithm can be stopped when no appreciable difference in sensitivity and specificity is observed for the last several boosting iterations. Note, however, that even after observed values of sensitivity and specificity in training data reach 100% the boosting classifier can improve sensitivity and specificity in testing data, a property of large margin classifiers (Schapire *et al.*, 1998).

5. RESULTS OF THE PROSTATE CANCER BIOMARKER-DISCOVERY PROJECT ANALYSIS

Two-stage classification of three pathologic states

The goal of our analysis strategy was to develop a classification rule that distinguishes the three pathologic states (cancer, BPH, and normal as defined) from serum samples based on their SELDI output: no attempt was made to separate early- vs. late-stage prostate cancer in this analysis. The three pathologic states were defined clinically by the screening results of PSA, digital rectal exam, and prostate biopsies, as described in Section 2. Because the 'BPH' group was defined by a *lack* of positive biopsies, it may contain some cancer patients if the biopsies missed small cancerous cells (see Section 6 for discussion of this issue). In addition, our simple exploratory analysis found close similarities in the peak/non-peak profiles between cancer and BPH samples in contrast to normal samples. We, therefore, decided to employ a two-stage approach to the classification of the three pathologic states. In the first stage, we seek a classification

rule that distinguishes cancer and BPH samples from normal samples. In the second stage, we attempt to separate cancer samples from BPH samples.

Pre-analysis processing

Through the pre-analysis processing of SELDI output (as described in Section 3), prior to the mass/charge axis alignment, we identified approximately 160–230 peaks in the 11 175 points of each SELDI output covering the mass/charge range of 1500–20 000. The range of the peak counts per sample, 160–230, was similar across the three pathologic states.

Construction of a summary classifier for cancer/BPH vs. normal classification by the boosting

Figure 4 shows the performance of the boosting summary classifiers for the cancer/BPH vs. normal classification in the training data (panel (a)) and in the test data (panel (b)) by the number of boosting iterations. Note that the boosting summary classifier was constructed using the training data and then applied to the test data. The performance measures are sensitivity (the proportion of cancer/BPH samples correctly classified) and specificity (the proportion of normal samples correctly classified). After 26 iterations, the classification error in the training dataset reached and remained at zero (Figure 4(a)). For a stopping rule of 100% sensitivity and specificity in the training data, for example, the final boosting summary classifier was s_{26} and its empirical sensitivity and specificity in the test data were 97.8% (44/45) and 100% (15/15), respectively (Figure 4(b)). A logistic regression with a forward variable-selection ($p < 0.01$) resulted in a similar classification performance: the difference in the area under the ROC curve was not large (0.996 for boosting vs. 0.967 for logistic) and was not statistically significant ($p = 0.14$).

Construction of a summary classifier for cancer vs. BPH classification by the boosting

Figure 5 shows the performance of the boosting summary classifiers for the cancer vs. BPH classification in the training data (panel (a)) and in the test data (panel (b)) by the number of boosting iteration. The performance measures are again sensitivity (the proportion of cancer samples correctly classified) and specificity (the proportion of BPH samples correctly classified). The classification error in the training dataset decreased as the number of iterations increased (Figure 5(a)), but the rate of decrease was slower than that for the classification of cancer/BPH vs. normal samples in Figure 4(a). The classification error was higher for the BPH samples than the cancer samples in both the training and test datasets. For a stopping rule of $\geqslant$90% sensitivity and specificity in the training data, for example, the final boosting summary classifier was reached at the 25th iteration, s_{25}, and its empirical sensitivity and specificity in the test data were 93.3% (28/30) and 46.7% (7/15), respectively (Figure 5(b)).

6. Discussion

Discovery of 'signature' profiles specific to each pathologic state and identification of individual proteins that form those signature profiles are key steps towards early detection of cancer. The method proposed here is intended for biomarker discovery based on high-dimensional mass spectrometry proteomic data. With the high dimensionality of the data, it is crucial to recognize the issue of overfitting in the search of biomarkers. It is easy to find some false profiles that fit the training dataset nearly perfectly with a large number of potential markers. To assess the degree of overfitting, we split samples into training and test sets and used the test set to provide independent estimates of classification errors for the classifiers derived from the training set. Alternatively, one may consider a cross-validation of the entire samples. An advantage of the cross-validation is that the entire set of available data is utilized in deriving classifiers,

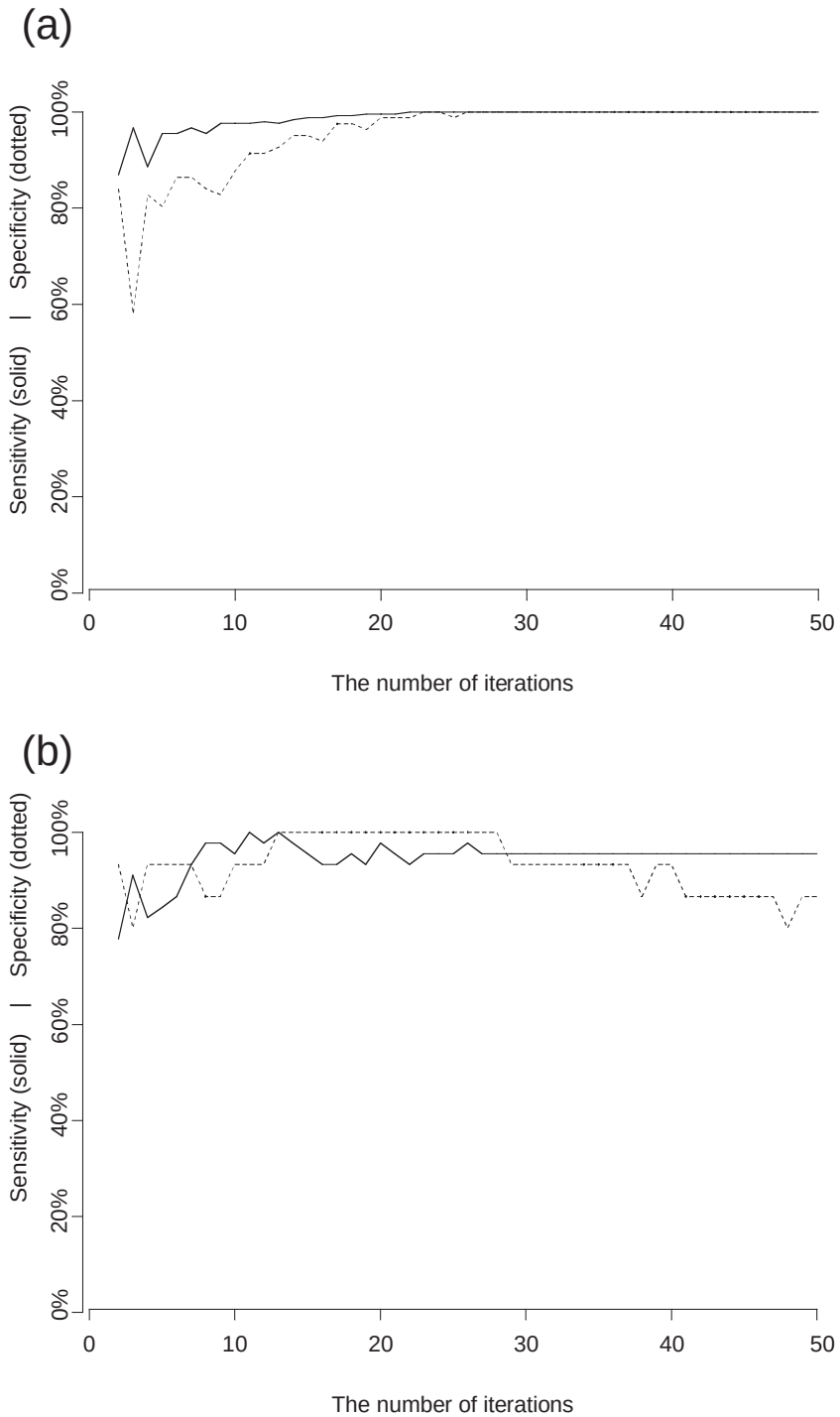

Fig. 4. Sensitivity (% cancer/BPH samples correctly classified) and specificity (% normal samples correctly classified) of the boosting summary classifier for the cancer/BPH vs. normal classification in the training dataset (a) and test dataset (b) by the number of boosting iteration.

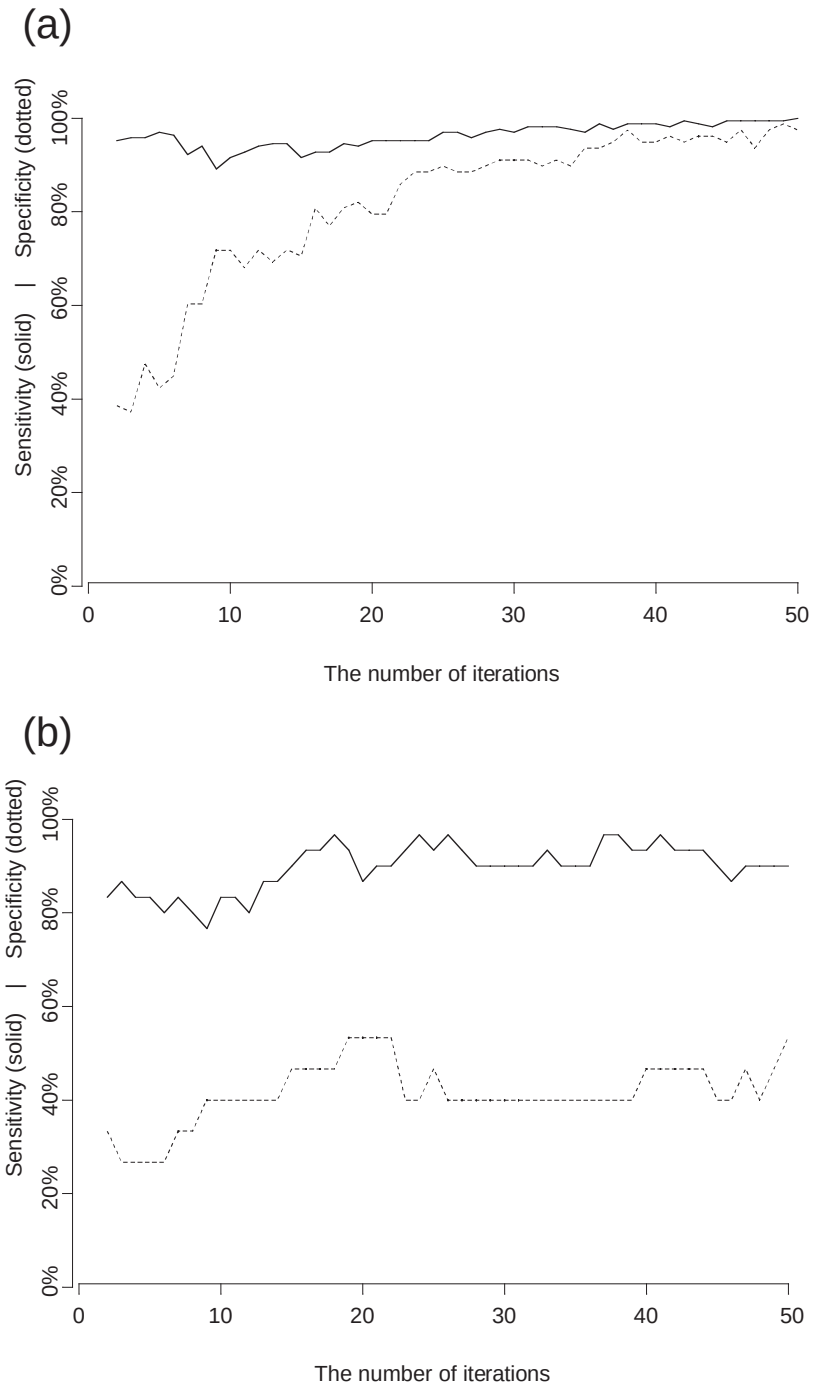

Fig. 5. Sensitivity (% cancer samples correctly classified) and specificity (% BPH samples correctly classified) of the boosting summary classifier for the cancer vs. BPH classification by the number of boosting iteration in the training dataset (a) and test dataset (b).

which is particularly relevant when the size of the data is small. A disadvantage is that, for a classifier selected by cross-validation, independent unbiased estimates of classification errors cannot be obtained.

Note that the lack of a positive biopsy for 'BPH' samples does not guarantee the absence of cancerous cells. For example, a large European cohort study reported that over 10% of men with PSA 4–10 ng/ml, who were negative by three biopsies (a transrectal ultrasound-guided sextant biopsy and two additional transition zone biopsies) were found to be positive by a repeat biopsy six weeks later (Djavan *et al.*, 2001). Thus, our 'BHP' group is likely to contain at least some cancer samples. Care must, therefore, be used in the interpretation of the results in the second-stage classification of cancer vs. BPH samples. Specifically, the mislabeling of cancer patients as 'BHP' would obscure the construction of the boosting summary classifier from the training data and perhaps lower the classification power of the resulting classifier. The large discrepancy between the specificity estimates in the training and test datasets (90% vs. 47%), but not in the sensitivity estimates (97% vs. 93%), is perhaps an indication of the mislabeling problem in the 'BHP' group. To evaluate this speculation, we examined the boosting weights of the 78 BPH and the 167 cancer samples in the training dataset at the 25th boosting iteration in the classification of cancer vs. BPH, where the boosting summary classifier reached $\geqslant$90% of both sensitivity and specificity in the training data (Figure 6). The distribution of the 78 BPH weights is appreciably wider than that of the 167 cancer weights, consistent with the notion that the 'BHP' group may be more heterogeneous than the cancer group, possibly due to the mislabeling problem. If the mislabeling problem was indeed present in the 'BHP' group, the estimated specificity of 47% in this group was an underestimate of the true specificity, which would be estimable only if we could exclude, with certainty, the cancer cases from the 'BHP' group. Future methodologic research needs to develop effective strategies for the classification problem where some of the true states are mislabeled.

Alternatively, the large discrepancy in the specificity between the training and test datasets suggests a possibility of overfitting, although the test dataset to estimate specificity was small (15 BPH samples). Nonetheless, the current screening method by PSA does not distinguish cancer and BPH well, and a biopsy, an invasive procedure, is required to separate cancer subjects from BPH subjects. Therefore, any new test that has a high sensitivity together with a specificity value appreciably different from 0% would be of interest clinically. Our cancer vs. BPH classifier meets this criterion with a very high sensitivity and a specificity approximately 50% in the test dataset.

To classify more than two states (e.g. the three pathologic states of cancer, BPH, and normal), multi-state boosting methods (e.g. Friedman *et al.*, 2000) can be used. In our analysis, however, they were not directly applicable because the 'BHP' group was likely to contain some cancer cases. Our two-stage approach was partly motivated by the potential mislabeling of the true state in the 'BHP' group: the first-stage (cancer/BPH vs. normal) classification is not influenced by the mislabeling of cancer as BPH.

The distribution of the three pathologic states (cancer, BPH, normal) in the test dataset was similar to that in the training dataset in our example. This was because we selected a simple random sample of 15 subjects from each of the four approximately equal-sized groups (early-stage cancer, late-stage cancer, BPH, normal) as the test dataset. It is important to note that, in a disease vs. non-disease classification, the prevalence of the disease in the training data influences the intercept of the boosting summary classifier: the boosting summary classifier calculates log odds of the disease for the classification which is a function of the disease prevalence. This has a practical implication: a boosting summary classifier built using a case-control dataset would be useful for ordering subjects in a screening population according to their estimated probability of being a disease subject, but the cutoff value of the log odds of the disease for defining positivity would have to be adjusted with an estimate of the disease prevalence in the screening population, if it is available. The main contribution of case-control studies in the biomarker discovery/validation process is, however, the identification of biomarkers that are likely to be useful for early detection of cancer (Pepe *et al.*, 2001): the exact formula for classification must be determined by subsequent longitudinal cohort studies.

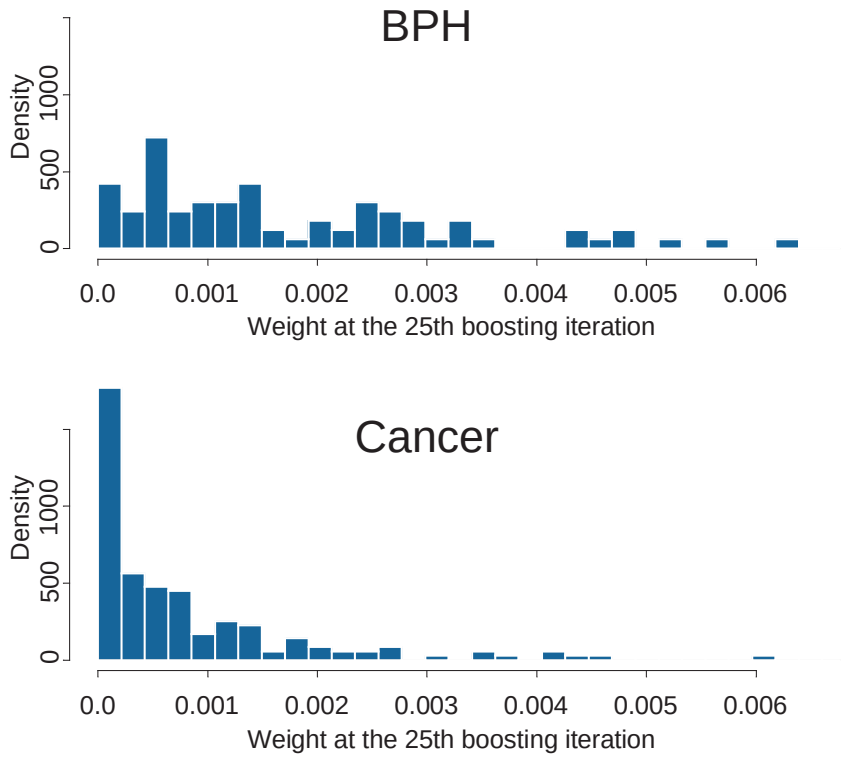

Fig. 6. Distribution of the boosting weights among the 78 BPH and the 167 cancer samples in the training dataset at the 25th iteration for the classification of cancer vs. BPH samples.

The biomarker discovery study discussed here is the first exploratory stage of biomarker discovery research and it is to be followed by studies with more rigorous designs (Pepe *et al.*, 2001). In order to identify true biomarkers, each subject's disease group (cancer, BPH, or normal) must be known without error. This requires a prostatectomy, or at least biopsies, for each subject, which is not viable under the current medical practice for subjects with normal PSA and digital rectal exam. Our study, as a first step of the biomarker discovery, assessed whether protein profiles can distinguish the three disease groups as defined by PSA values and biopsy results triggered by elevated PSA values. We are currently conducting a second study, a multi-center project, in which we pool normal and BPH subjects together and concentrate on the classification of cancer vs. non-cancer groups and aggressive vs. non-aggressive cancer types. The third step is planned to utilize large cohort samples from a randomized trial of chemoprevention of prostate cancer in which a biopsy is taken from each subject regardless of PSA values: the trial is currently underway.

With further improvements in the measurement properties of the SELDI technology, it will become more reasonable to use intensity measures as continuous variables, instead of reducing them to binary peak/non-peak indicators as we did here. Also, the alignment of mass/charge axis in the analysis stage may become unnecessary in the future. Technological enhancements of SELDI towards these goals are under active developments in laboratories including ours at the Eastern Virginia Medical School and Ciphergen Biosystems Inc. Under the current measurement properties of the SELDI technology, however, innovative analytic strategies such as those proposed here are needed for productive searches of protein biomarkers.

As a yardstick, a standard classification method without any pre-analysis processing of SELDI output (i.e. a logistic discriminant analysis that used a stepwise selection with $p < 0.05$ of mass/charge points with the original intensity measures as predictors) resulted in the following mediocre performances in the test dataset for the classification of cancer vs. normal samples, the easiest classification among the three pathologic states: (sensitivity, specificity) of (86.7%, 73.3%), (83.3%, 86.7%), and (73.3%, 86.7%) with the cutoff at the predicted probability of 0.5, 0.7, and 0.9, respectively. Improvements of the proposed method may be possible using, instead of the boosting, other modern classification approaches such as support vector machines (Vapnik, 1998), which are closely related to the boosting (Rätsch *et al.*, 2000; Freund and Schapire, 1999), or logic regression (Ruczinski *et al.*, to appear). It is our hope that the data-analytic strategy proposed here will stimulate further advances and alternative approaches to the disease-state profiling based on high-dimensional proteomic data and contribute to the discovery of useful biomarkers.

ACKNOWLEDGEMENTS

This research was supported by Grant U01-CA86368 from the National Cancer Institute and Grant #DAMD17-02-1-0054 from the Department of Defense Prostate Cancer Research Program (PCRP) of the US Army Medical Research and Materiel Command's Office of the Congressionally Directed Medical Research Programs (CDMRP).

REFERENCES

ADAM, B.-L., VLAHOU, A., SEMMES, O. J. AND WRIGHT, JR., G. L. (2001). Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics* **1**, 1264–1270.

ALBERTS, B., BRAY, D., LEWIS, J., RA, M., ROBERTS, K. AND WATSON, J. D. (1994). Molecular Biology of the Cell (3rd ed.). New York: Garland.

DJAVAN, B., MAZAL, P., ZLOTTA, A., WAMMACK, R., RAVERY, V., REMZI, M., SUSANI, M., BORKOWSKI, A., HRUBY, S., BOCCON-GIBOD, L., SCHULMAN, C. C. AND MARBERGER, M. (2001). Pathological features of prostate cancer detected on initial and repeat prostate biopsy: results of the prospective European Prostate Cancer Detection study. *Prostate* **47**, 111–117.

FREUND, Y. (1995). Boosting a weak learning algorithm by majority. *Information and Computation* **121**, 256–285.

FREUND, Y. AND SCHAPIRE, R. E. (1996). Experiments with a New Boosting Algorithm. *Machine Learning: Proceedings of the Thirteenth International Conference*. pp. 148–156.

FREUND, Y. AND SCHAPIRE, R. E. (1999). A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence* **14**, 771–780.

FRIEDMAN, J. H. (1984). *A Variable Span Smoother,* Technical Report No. 5. Laboratory for Computational Statistics, Dept. of Statistics, Stanford University, CA.

FRIEDMAN, J. H., HASTIE, T. AND TIBSHIRANI, R. (2000). Additive logistic regression: a statistical View of Boosting. *Annals of Statistics* **28**, 337–407.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2001). *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. New York: Springer.

LANDER, E. S. (1999). Array of hope. *Nature Genetics* **1 Supplement**, 3–4.

LIOTTA, L. AND PETRICOIN, E. (2000). Molecular profiling of human cancer. *Nature Reviews Genetics* **1**, 48–56.

PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THORNQUIST, M., WINGET, M. D.

AND YASUI, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 1054–1061.

RÄTSCH, G., SCHÖLKOPF, B., MIKA, S. AND MÜLLER, K. R. (2000). *SVM and Boosting: One class*, Technical Report 119. Berlin: GMD FIRST.

RIDGEWAY, G. (2000). Discussion of 'Additive Logistic Regression: a Statistical View of Boosting' by Friedman, J., Hastie, T., and Tibshirani, R.. *Annals of Statistics* **28**, 393–400.

RUBIN, R. B. AND MERCHANT, M. (2000). A rapid protein profiling system that speeds study of cancer and other diseases. *American Clinical Laboratory* **19**, 28–29.

RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. L. Logic Regression. *Journal of Computational and Graphical Statistics* (to appear).

SCHAPIRE, R. E. (1990). The strength of weak learnability. *Machine Learning* **5**, 197–227.

SCHAPIRE, R. E., FREUND, Y., BARTLETT, P. AND LEE, W. A. (1998). Boosting the margin: A new explanation for the effectiveness of voting methods. *The Annals of Statistics* **26**, 1651–1686.

SRINIVAS, P. R., SRIVASTAVA, S., HANASH, S. AND WRIGHT, JR., G. L. (2001). Proteomics in early detection of cancer. *Clinical Chemistry* **47**, 1901–1911.

SRIVASTAVA, S. AND KRAMER, B. S. (2000). Early detection cancer research network. *Laboratory Investigation* **80**, 1147–1148.

VAPNIK, V. (1998). *Statistical Learning Theory*. New York: Wiley.

WRIGHT, JR., G. L., CAZARES, L. H., LEUNG, S.-M., NASIM, S., ADAM, B.-L., YIP, T.-T., SCHELLHAMMER, P. F., GONG, L. AND VLAHOU, A. (1999). Proteinchip surface enhanced laser desorption/ionization (SELDI) mass spectrometry: A novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostate Disease* **2**, 264–276.

[*Received July 17, 2002; revised November 14, 2002; accepted for publication December 18, 2002*]