Computational Protein Biomarker Prediction: a Case Study for Prostate Cancer

D. NAIK, A. POTHEN, M. WAGNER, S. KASUKURTI, R. DEVINENI, B-L. ADAM, O. J. SEMMES, and G. L. WRIGHT Jr.

ABSTRACT

Recent technological advances in mass spectrometry now pose challenges in computational mathematics and statistics to process the mass spectral data into predictive models with clinical and biological significance.

We discuss computational approaches to find protein biomarker candidates using protein profiles obtained via mass spectrometry and assess their statistical significance. Our overall goal is to implicate peaks that have a high likelihood of being affected by a given disease state, and thus to narrow the search for biomarker candidates. Thorough cross-validation studies and randomization tests are performed on a prostate cancer dataset with over 300 patients, obtained at the Eastern Virginia Medical School using SELDI-TOF mass spectrometry. Average classification accuracy of 87% on a 4-group classification problem was obtained using a two-stage SVM-based procedure and just 16 peaks, with other methods performing comparably.

Keywords: ANOVA F-statistic, B/W ratio, biomarker discovery, feature (protein) selection, prostate cancer data, SELDI-TOF mass spectrometry, statistical discrimination methods, support vector machine.

1 Department of Mathematics & Statistics, Old Dominion University, Norfolk, Virginia 23529. E-mail: dnaik@odu.edu
2 Department of Computer Science, Old Dominion University, Norfolk, Virginia 23529. E-mail: {pothen, skasukur, devin_r}@cs.odu.edu
3 Division of Pediatric Informatics, Cincinnati Children’s Hospital Medical Center, Cincinnati, OH 45229. E-mail: mwagner@chmcc.org
4 Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, Virginia 23507. E-mail: {adambl, semmesoj, wrightgl}@evms.edu
1 Introduction

Recent advances in mass spectrometry (MS) technology are starting to enable high-throughput profiling of the protein content of complex samples. It is foreseeable that MS, coupled with chromatographic separation techniques, might become complementary to microarray technology on the proteome level. The very dynamic nature of the proteome, the wide range of abundances, the general lack of a “protein catalogue” (unlike the genomic catalogue, which is all but complete for a number of organisms) and various technical challenges in capturing proteins make this a particularly ambitious and challenging undertaking. While MS has been used extensively on purified, digested samples to identify proteins via peptide mass fingerprints, the data we use in this paper is fundamentally different since it consists of mass spectra (or, more precisely, peaks from mass spectra) of complex mixtures such as blood serum. After some chromatographic separation steps (which are crucially important, but not the primary subject of this paper), a mass spectrum of the matrix-crystallized sample is obtained on a wide mass range (in our case, 2-40 kDa) in order to obtain a profile of the protein content of a sample. If reproducibility is ensured, then these spectra can be used to identify peaks whose intensities correlate with a particular phenotype of interest, e.g., in this paper, prostate cancer.

The purpose of this paper is to show that computational methods can be useful in narrowing the search for protein biomarker candidates. Once we find a small set of peaks that can be used to computationally “predict” phenotypes with high accuracy, then these peaks should be analyzed further and the underlying proteins identified, e.g., by focusing an MS/MS instrument on the relevant peak masses. The hope is that the subsequent functional study of these proteins will eventually lead to new biological insights into disease pathways and, ultimately, to reliable diagnostic tests and potential therapeutic targets. We want to stress the need for biological validation; the inherent variability of mass spectrometry data makes it uncertain whether peak profiles can be used for diagnosis directly.

The need for computational methods is evident in order to find peaks that correlate with phenotypes and, equally importantly, in order to assess their statistical significance. We survey several classification (or supervised learning) methods that can be used in this context and apply them to a multi-class prostate cancer prediction problem. Given that sample sizes for these kinds of experiments are typically small, and given that validation of any results we produce requires laborious protein identification, our aim is to find the smallest set of peaks that yields reasonable (i.e., statistically significant) classification results.

Various statistical and other optimization techniques for classification are described in the literature. For a recent and excellent review of these methods see Hastie, Tibshirani, and Friedman (2001). Some of these methods have been successfully utilized for discriminating cancer samples from normal samples. For example, Petricoin et al. (2002) used genetic algorithms for feature selection and classification in their study of ovarian cancer, Adam et al. (2002) and Qu et al. (2002) used the “area under curve (AUC)” criterion for feature selection and decision trees in conjunction with boosting techniques for classification in their prostate cancer study. Li et al. (2002) used the signal to noise ratio for an initial feature selection and, subsequently, used the unified maximum separability analysis (UMSA) algorithm repeatedly for classification in their breast cancer study.

The data set we analyze in the article was obtained at the Virginia Prostate Center of the Eastern Virginia Medical School using SELDI-TOF (surface enhanced laser desorption and ionization—
time of flight) mass spectrometry. Qu et al. (2002) and Adam et al. (2002) have used data from the same source to build hierarchical decision trees; here we suggest improvements by using and comparing several standard optimization- and statistics-based classification techniques and by assessing prediction significance via randomization techniques.

By dividing the data set at our disposal into training set (to be used for model building) and test set (to estimate the generalization power of the model), and by doing this randomly and many times over, we can benchmark various classification methods reliably and gain insights into their capabilities of handling proteomic data. In particular, we used Fisher’s linear and quadratic discriminant functions, nonparametric kernels, nearest neighbor methods and linear support vector machines for classification. While performing the cross-validation studies, care must be taken so that the test set does not influence the choice of the peaks used in the classification. Misclassification rates are biased (downwardly) when both the training and test sets are used for feature selection as opposed to when only the training set is used.

Although we will use the above mentioned prostate cancer data for illustration, the data analysis strategies and the methods used here are of general applicability and could easily be adapted for other mass-spectrometry datasets. In earlier work (Wagner, Naik and Pothen, 2003) we have proposed candidate biomarkers in lung cancer using protein profiles obtained via MALDI-TOF (matrix assisted laser desorption and ionization–time of flight) mass spectra (data provided by Duke University’s department of Radiology) by employing similar data analysis strategies and classification methods.

2 Materials and Methods

2.1 Samples

Serum samples were obtained from the Virginia Prostate Center Tissue and Body Fluid Bank. Surface enhanced laser desorption/ionization time-of-flight (SELDI-TOF) mass spectrometry protein profiles of serum from 82 unaffected healthy men, 77 patients diagnosed with benign prostatic hyperplasia (BPH), 83 patients with organ-confined prostate cancer (PCA), and 82 patients with non-organ-confined PCA were available, in duplicate, for the analysis. For details on sample preparation and the particular kind of chromatographic affinity chip used, see Adam et al. (2002).

Each spectrum is an array of intensities of the signal at discretely sampled values of the mass to charge ratio of the ions. A sample mass spectrum (from a different data set) is shown in Figure 1. At greater resolution, one can see much more “structure” to each peak.

Mass spectra of protein mixtures extracted from samples contain, in general, in the order of several hundreds of candidate peaks along the mass (m/z) axis for each sample, and the data set contains the masses and the intensities of the peaks at these mass values for each spectrum. The number of peaks varies from sample to sample and, furthermore, the same peak may occur at different mass locations in different samples. Hence peak detection and alignment of the peak become crucial issues in the analysis. This process involves baseline subtraction, peak identification and extraction, intensity normalization, and peak alignment across samples.

Peak detection and alignment of our data were performed with Ciphergen ProteinChip Software 3.0 with some modifications. All 652 spectra (326 samples with replicates) were compiled and 779 peaks in the mass range from 2 to 40kDa were selected by the ProteinChip software for analysis.
Figure 1: Sample protein mass-spectrum with baseline identified.
This range contains the majority of the resolved protein/peptides (Qu et al., 2002). Details of the steps involved in this pre-processing of the data are given in Adam et al. (2002).

2.2 Peak Selection

Feature selection, i.e., the reduction of the number of input variables (or, in our case, peaks), is a crucially important step. Many classification methods are known to perform poorly when “irrelevant” features or ones without information content are added. Secondly, computational biologists are frequently faced with the problem of having only few (tens) samples but many (thousands) descriptors, as is the case with microarray analysis. This presents the challenge of designing models that are not “overfitted” to the data. One approach to prevent this is to try to decrease the feature dimensionality by performing feature selection.

In our case, we are interested in finding a reasonably small set of peaks in order to then enable the identification of the underlying proteins and, eventually, understand the biological function they have in the disease pathway. In this sense the classification methods used can be viewed as validation methods for the feature selection algorithms.

Unfortunately, finding the “best” set of features to build a predictive model is a hard combinatorial problem, and so one must live with heuristic approaches. The literature on this subject is vast, and one generally distinguishes between filtering methods (those which rank individual features according to some criterion) and more involved wrapper algorithms, which use classification methods directly to evaluate a particular set of features.

For this paper we use only simple filter methods since they seem to do reasonably well for our purposes. We first chose to disregard any peaks which appear in 30 or fewer samples, thus preventing the classification methods from taking advantage of what are likely to be spurious peaks or data artifacts, possibly contaminants. In our particular case this resulted in a reduction from 779 peaks in the original dataset to 220. With over 300 samples at our disposal this was already a good starting point.

In order to further reduce the number of features to, say, under 25, we used the ratio of between group sum of squares and within group sum of squares (B/W ratio), which is equivalent to using the ANOVA F-statistic, for feature selection. For every variable in the data set (that is, for every mass value), we compute the ratio of between group sum of squares to within group sum of squares and select the data corresponding to the $q$ variables with the highest ratios. Suppose for the $j$th variable the between group sum of squares is $B_j = \sum_{i=1}^{g} (\bar{y}_{ij} - \bar{y}_j)^2$ and the within group sum of squares is $W_j = \sum_{i=1}^{g} \sum_{k=1}^{n_i} (y_{ikj} - \bar{y}_{ij})^2$, where $g$ is the number of groups, $n_i$ is the number of samples in the $i$th group, $\bar{y}_{ij} = \frac{1}{n_i} \sum_{k=1}^{n_i} y_{ikj}$, $\bar{y}_j = \frac{1}{\sum_{i=1}^{g} n_i} \sum_{i=1}^{g} \sum_{k=1}^{n_i} y_{ikj}$, and $y_{ikj}$ is the observed value of the $j$th variable of the $k$th sample belonging to the $i$th group.

For every feature $j = 1, \ldots, p$ we compute $B_j/W_j$ or equivalently, the ANOVA F-statistic $F_j = \frac{B_j/\nu_1}{W_j/\nu_2}$, where $\nu_1 = g - 1$ and $\nu_2 = \sum n_i - g$ are the degrees of freedoms of $B_j$ and $W_j$ respectively. Then the reduced data set will be the data corresponding to the $q$ largest $F_j$ values.

An advantage of this method over the other data reduction techniques, like principal component analysis, where the data on few linear combinations of the original variables are used, is that we can indeed identify the important proteins (by their molecular weight or mass-to-charge values) that are used for the analysis of the data.

A slightly more general method for feature selection is to use the Wilks’ likelihood ratio ($\lambda$)
which is the ratio of maximum likelihood function under the assumption that there is no difference between the groups to maximum likelihood function under no such assumption. Smaller values of $\lambda$ indicate more significant difference between the groups. If the probability density functions ($pdf$) are normal then this method reduces to the B/W ratio (F-statistic) method described above. The advantage of this method is that it allows us to use different $pdf$ for this process.

### 2.3 Classification/Discrimination Methods

The next task is to perform a discriminant analysis to construct discriminant functions so that the classification of the new unknown samples obtained from MS can be performed. Various classical and modern methods are available for this purpose. Classical statistical methods (parametric as well as nonparametric) have stood the test of time and proved to be very useful methods. However, two modern classification methods have emerged recently. One set of methods is bagging with boosting of classification trees and the other set is based on support vector machines. Boosting methods have been utilized by Qu et al. (2002). Here we will adopt several classical statistical methods and support vector machines for our analysis. In the following we will briefly describe these methods; more details can be found in Khattree and Naik (2000) and Hastie, Tibshirani, and Friedman (2001).

**Likelihood based discriminant procedures:** In general, in discriminant analysis, the decision rule to classify a new (or test) sample into one of the several groups, taking the prior probabilities and the cost of misclassifications into consideration, is as follows. Classify the sample with an observation vector $y$ into the $i$th group if the expected cost of misclassification, $\sum_{s=1}^{g} \pi_s f_s(y) c(i | s), s \neq i$, is smallest for $i = 1, ..., g$. Here $\pi_s$ and $f_s(y), s = 1, ..., g$, are respectively the prior probability and the probability density function for the $s$th group and $c(i | s)$ is the cost of misclassification when the sample is classified into the $i$th group when it actually comes from the $s$th group. Of course, for $s = i$, $c(i|i) = 0$. If all the costs of misclassification are assumed to be equal then the classification rule is based on minimizing the expected total probability of misclassification and we classify the sample into the $i$th group if $\sum_{s=1}^{g} \pi_s f_s(y), s \neq i$, is smallest for $i = 1, ..., g$. This rule further reduces to simply checking whether or not $\pi_i f_i(y) > \pi_j f_j(y)$, for all $j = 1, ..., g, j \neq i$.

In practice, different known multivariate probability densities can be used for $f_i(.)$, but the most common density used is the multivariate normal density. If the form of the density is not assumed to be known then the nonparametric methods are used for estimating the density using the data. If the probability density function for the $i$th group is assumed to be multivariate normal with mean vector $\mu_i$ and variance covariance matrix $\Sigma_i$ for $i = 1, ..., g$, then the above classification rule leads to the *quadratic discrimination rule*; it reduces to the *linear discrimination rule* if the variance covariance matrices $\Sigma_i's$ are all equal for the $g$ groups. If the form of the $pdf$ is not known then the data are used to estimate the $pdf$, generally using the *kernel method*.

**Fisher’s canonical discriminant analysis** provides a method where no probability density functions are directly used, instead only a few (less than or equal to $g - 1$) canonical variables, which are certain linear combinations of the original variables, are employed. The canonical variables have the better capacity to discriminate between the groups than any individual variable because these are created such that the between group sum of squares for these variables is larger relative to the within group sum of squares.
The nearest neighbor method provides another approach which is based on distances from ‘immediate neighbors’ and hence bypasses the need for a probability density. Two common distance measures used in this context are, (i) the Mahalanobis distance or Euclidean distance and (ii) one minus the absolute value of the correlation coefficient, between the two samples. In this method we first compute the affinity measure between the unknown sample and all the other samples. There will be as many values of the affinity measure as there are the number of samples. Next, we find \( k \), a pre-specified number, samples corresponding to the \( k \) smallest values of the selected affinity measure. We classify the unknown sample to the group based on a simple majority rule. It is clear that we may have undecided cases in this approach.

Support vector machines (SVMs) are powerful classification tools that arose out of the machine learning and optimization communities in the 1960’s (e.g. Mangasarian, 1965). If the data is separable, linear SVMs compute a separating hyperplane that maximizes the distance (the margin) to any training vector. In the nonseparable case SVMs find a separator that optimizes an objective function comprised of a weighted sum of misclassification and margin size, where the tradeoff between the two terms is controlled explicitly by a parameter. More concisely, for given training points \( y_i \) in groups \( C_0 \) and \( C_1 \), linear SVMs find the hyperplane \((w, b)\) that solves the quadratic optimization problem

\[
\min_{w, b, \xi} \|w\|_2 + C\|\xi\|_1 \\
\text{such that } a_i(y_i'w + b) + \xi_i \geq 1 \text{ for all } i \in C_0 \cup C_1.
\]

Here \( w \) is the normal vector that together with the displacement \( b \) defines the separating hyperplane, \( a_i = 1 \) for \( i \in C_1 \) and \( a_i = -1 \) for \( i \in C_0 \), \( \xi = (\xi_i) \) is the vector of so-called slack variables, and \( C \) is a tradeoff parameter.

The idea can be extended to nonlinear decision boundaries by introducing nonlinear kernels. By choosing appropriate kernel functions \( h(y_i) \) that transform the data point, one can define support vector machines that handle polynomial or Gaussian kernels. The reader is referred to Cristianini and Shawe-Taylor (2000) for details.

The extension of SVMs to the case with multiple classes such as our particular application is still an active research topic. Lee, Lin and Wahba (2001) have found natural and theoretically satisfying extensions, however, for simplicity we have opted for another scheme which, despite its simplicity, has shown quite satisfactory results in practice. We opted to adapt the pairwise approach that constructs all \( n(n - 1)/2 \) pairwise discriminators for \( n \) classes, an approach that is reasonable for small values of \( n \). The final classifier is taken to be the one that dominates all others, if one exists. Otherwise the result is considered to be inconclusive, an event which occurs in only a very small percentage of cases. We want to stress here that even in the inconclusive cases it is sometimes possible to rule out certain classes (in case they are dominated by all others), which is an outcome that might still be of some medical relevance.

## 3 Results and Discussion

In order to assess the generalization power of the classification methods and to estimate their prediction capabilities for unknown samples, we split the data into training and test sets. To make
# of peaks used

|        | 10 | 15 | 20 | 25 | 30 | 35 | 50 | 70 |
|--------|----|----|----|----|----|----|----|----|
| Quadr. Discr. | 74.3 | 74.4 | 73.2 | 74.7 | 76.9 | 77.3 | 80.1 | 76.3 |
| Nonpar (Kernel) | 75.9 | 77.2 | 76.2 | 78.1 | 78.5 | 79.3 | 79.9 | 75.4 |
| kNN | 74.8 | 75.9 | 75.7 | 77.5 | 75.5 | 76.3 | 73.3 | 70.7 |
| Fisher Linear | 72.6 | 77.6 | 80.8 | 80.9 | 81.9 | 85.0 | 85.3 | 84.5 |
| SVM | 74.7 | 78.7 | 81.5 | 80.8 | 80.7 | 80.8 | 82.8 | 82.2 |

Table 1: Classification accuracy (in percent) of various classification methods on the full 4-class prostate cancer dataset using various numbers of peaks.

the results comparable for each of the following results, the training sets consisted of randomly chosen subsets containing 90% of each class. The remaining 10% of the samples from each class were left as test sets. We stress that feature selection was performed in every experiment on the training set only (unlike what is often seen in the literature) in order not to bias the feature selection procedure unfairly. Several papers (including Wagner, Naik and Pothen, 2003) have shown that performing feature selection on the entire dataset often grossly underestimates the generalization error.

In what follows and unless otherwise indicated, all error rates are averaged over 100 runs. Table 1 reports experiments on the original dataset with samples categorized into 4 groups: normal, BPH, early (localized) cancer and late (metastasized) cancer. Results for the first four methods reported were obtained with codes implemented in SAS, results for the linear SVM were obtained with the package SvmFu (which is freely available at the website http://five-percent-nation.mit.edu/SvmFu/). We made initial choices of the parameters in the classification methods before doing the experiments, and did not tune them for optimal classification of this data set (e.g., the “trade-off” parameter \( C \) in SVM was set at one for all runs).

|        | BPH | Late Cancer | Early Cancer | Normal |
|--------|-----|-------------|--------------|--------|
| BPH    | 752 (94.0%) | 47 (5.9%) | 0 (0%) | 1 (0.1%) |
| Clinical | Late Cancer | 89 (11.1%) | 570 (71.3%) | 128 (16.0%) | 13 (1.6%) |
| Diagnosis | Early Cancer | 36 (4.5%) | 94 (11.8%) | 656 (82.0%) | 14 (1.8%) |
| Normal | 24 (3.0%) | 19 (2.4%) | 15 (1.9%) | 742 (92.8%) |

Table 2: Details of classification results obtained with Fisher’s Linear Discriminator and 35 peaks on the full 4-class problem. The overall average classification accuracy is 85.0%.

As can be seen in Table 1, the methods achieve rather comparable prediction accuracies, with the best cross–validated result being obtained in this case by the linear discriminators. These results should be viewed in the context of what one would expect to see if the peaks considered contained no information with regards to the various phenotypes. Since there are 4 classes, a random classifier would be expected to achieve about 25% accuracy.
In order to get a sense of the significance of these results and to attempt to rule out data artifacts, we checked the performance of the classifiers on the same data but with randomized group assignments. We generated 1000 randomized datasets and averaged the performance of the SVM using 15 peaks on 10 random choices of test and training set (so that in fact 10,000 random runs were performed). The best classification accuracy average out of those 1000 runs was 34.4%, while the median classification accuracy was 24.1%. This is significantly below the 78.7% reported in Table 1 and is an indication that these results are not merely due to some structure in the data.

Finally, Table 1 also illustrates that all methods are rather sensitive to noise. Increasing the number of peaks at times deteriorates the classification accuracy, underscoring the need for high-quality feature selection procedures. As mentioned in the introduction, our aim is to find a small set of peaks that have good prediction capabilities. The results presented here are meant to assess the generalization capabilities of the modeling approach, the “final” set of peaks can then of course be chosen using the entire set. Conclusions to be drawn from the particular peaks here is the subject of future research.

For illustration purposes we show detailed results obtained with Fisher’s Linear Discriminator on the full 4-class problem in Table 2. As mentioned above, 10% (8 samples) of each class were left as test set and 100 runs were performed. We note that by far the largest source of misclassification is the distinction between early and late cancers, maybe indicating that this distinction is also one which is difficult clinically. In any case, we want to stress again that our aim is not so much to achieve perfect classification but rather to gather evidence that at least some of the underlying peaks are likely to be implicated in the disease. We believe that this goal has been achieved.

|               | # of peaks used (malignant vs. other) |
|---------------|--------------------------------------|
|               | 5 | 8 | 10 | 12 | 15 |
| Quadr. Disc.  | 83.8 | 85.9 | 85.7 | 85.8 | 85.7 |
| Nonpar. (Kernel) | 84.2 | 88.2 | 88.0 | 88.4 | 87.3 |
| kNN           | 89.7 | 89.8 | 87.4 | 87.7 | 87.3 |
| Fisher Linear | 87.5 | 89.6 | 88.8 | 87.9 | 88.0 |
| SVM           | 88.6 | 90.7 | 90.9 | 91.9 | 90.9 |

Table 3: Classification accuracy on the data obtained by grouping all Normal and BPH samples into one class, and all cancer samples into another (Class sizes thus remain approximately balanced).

It turns out that we can further reduce the number of peaks required to classify accurately by considering a two-stage hierarchical classification procedure. First, we aim to distinguish whether a sample is benign (normal or BPH) or cancerous. As seen in Table 3, this can be done with high accuracy (90.7%) with only 8 peaks using an SVM. Table 4 shows the average prediction accuracy achieved on other pairwise discriminations, indicating that the normal versus BPH distinction can be made with 96% accuracy again using just 8 peaks. Thus we obtain at least 87% accuracy for the two-stage process with a total of 16 peaks, assuming we do not need to distinguish between early and late cancers. This procedure also implies an alternate feature selection strategy for multi-class problems: Instead of ranking features using the F-statistic criterion on the entire data set, choose the union of top-ranking features that score highest in pairwise comparisons.
Table 4: SVM classification average accuracy results for other pairwise distinctions using varying numbers of peaks.

| # of peaks used | 5  | 8  | 10 | 12 | 15 |
|-----------------|----|----|----|----|----|
| BPH vs Normal   | 94.9 | 96.7 | 96.4 | 95.9 | 95.7 |
| BPH vs E. Cancer | 94.7 | 94.2 | 94.3 | 94.9 | 95.2 |
| BPH vs L. Cancer | 89.1 | 88.1 | 88.9 | 89.7 | 91.7 |
| Normal vs E. Cancer | 86.8 | 92.9 | 95.2 | 94.4 | 95.1 |
| Normal vs L. Cancer | 88.0 | 88.7 | 88.5 | 90.5 | 90.4 |

Table 5: Statistics on classification accuracy for SVM averaged over 1000 randomized datasets, 10 runs on each dataset, using 15 peaks.

|               | max. acc. | median. acc | 95th %ile |
|---------------|-----------|-------------|-----------|
| BPH vs Normal | 70.0      | 51.6        | 59.7      |
| BPH vs E. Cancer | 68.1    | 50.0        | 59.4      |
| BPH vs L. Cancer | 68.1    | 50.0        | 59.7      |
| Normal vs E. Cancer | 66.9    | 50.0        | 59.7      |
| Normal vs L. Cancer | 65.0    | 51.6        | 59.3      |

Table 5 shows that the accuracy rates achieved in Table 4 are far better than any results on randomized data, giving us additional confidence that the chosen peaks are indeed significant.

References

B-L. Adam, Y. Qu, J.W. Davis, M.D. Ward, M.A. Clements, L.H. Cazares et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Research*, 62: 3609-3614, 2002.

N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*. Cambridge University Press, Cambridge, UK, 2000.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics, 2001.
R. Khattree and D. Naik. *Multivariate Data Reduction and Discrimination with SAS Software*. SAS Institute and J. Wiley and Sons, 2000.

Y. Lee and Y. Lin and G. Wahba. Multicategory support vector machines. Department of Statistics, University of Wisconsin, Technical Report-1043, 2001. *To appear, Proceedings of the 33rd Symposium on the Interface*, 2001.

J. Li, Z. Zhang, J. Rosenzweig, Y.Y. Wang and D.W. Chan. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clinical Chemistry*, 48: 1296-1304, 2002.

O. L. Mangasarian. Linear and nonlinear separation of patterns by linear programming. *Operations Research*, 13:444–452, 1965.

E. F. Petricoin III, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro et al. Use of proteomic patterns in serum to identify ovarian cancer. *The Lancet*, 359: 572-577, 2002.

Y. Qu, B-L. Adam, Y. Yasui, M.D. Ward, L.H. Cazares et al. Boosted decision tree analysis of SELDI mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clinical Chemistry*, 48: 1835-1843, 2002.

M. Wagner, D. Naik, and A. Pothen. Protocols for disease classification from mass spectrometry data. To appear in *Proteomics*, 2003.

G. L. Wright Jr., L.H. Cazares, S-M. Leung, S. Nasim, B-L. Adam et al. Proteinchip SELDI mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer and Prostatic Diseases*, 2:264–276, 1999.