BMC Bioinformatics

Research article

Computational protein biomarker prediction: a case study for prostate cancer

Michael Wagner*[1], Dayanand N Naik[2], Alex Pothen[3], Srinivas Kasukurti[3], Raghu Ram Devineni[3], Bao-Ling Adam[4], O John Semmes[4] and George L Wright Jr[4]

Address: [1]Cincinnati Children's Hospital Research Foundation and Department of Biomedical Engineering, University of Cincinnati, Cincinnati, OH 45229, USA, [2]Department of Mathematics and Statistics, Old Dominion University, Norfolk, VA 23529, USA, [3]Department of Computer Science, Old Dominion University, Norfolk, VA 23529, USA and [4]Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23507, USA

Email: Michael Wagner* - mwagner@cchmc.org; Dayanand N Naik - dnaik@odu.edu; Alex Pothen - pothen@cs.odu.edu; Srinivas Kasukurti - skasukur@cs.odu.edu; Raghu Ram Devineni - devin_r@cs.odu.edu; Bao-Ling Adam - adambl@evms.edu; O John Semmes - semmesoj@evms.edu; George L Wright - wrightgl@evms.edu

* Corresponding author

Published: 11 March 2004

BMC Bioinformatics 2004, **5**:26

Received: 06 December 2003
Accepted: 11 March 2004

This article is available from: http://www.biomedcentral.com/1471-2105/5/26

Abstract

Background: Recent technological advances in mass spectrometry pose challenges in computational mathematics and statistics to process the mass spectral data into predictive models with clinical and biological significance. We discuss several classification-based approaches to finding protein biomarker candidates using protein profiles obtained via mass spectrometry, and we assess their statistical significance. Our overall goal is to implicate peaks that have a high likelihood of being biologically linked to a given disease state, and thus to narrow the search for biomarker candidates.

Results: 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. We obtain average classification accuracies of 87% on a four-group classification problem using a two-stage linear SVM-based procedure and just 13 peaks, with other methods performing comparably.

Conclusions: Modern feature selection and classification methods are powerful techniques for both the identification of biomarker candidates and the related problem of building predictive models from protein mass spectrometric profiles. Cross-validation and randomization are essential tools that must be performed carefully in order not to bias the results unfairly. However, only a biological validation and identification of the underlying proteins will ultimately confirm the actual value and power of any computational predictions.

Background

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 catalog" (unlike the genomic catalog, 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 are fundamentally different since they consist 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, 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.

The literature on classification techniques as well as the related field of machine learning is vast; we will not even attempt to give a summary but will rather refer the readers to, e.g., the excellent book by Hastie, Tibshirani and Friedman [1]. Several of these methods have been utilized for discriminating cancer samples from normal (control) samples using proteomic data. The two main components of these approaches are 1) the feature selection (dimension reduction) method and 2) a classification method to build a predictive model. For example, Petricoin et al. [2] used genetic algorithms for feature selection and Kohonen's self-organizing map classification in their study of ovarian cancer (see also Sorace and Zhan [3] for a more detailed and principled analysis of the same data). Li et al. [4] used the signal-to-noise ratio for an initial feature selection and, subsequently, used "unified maximum separability analysis" repeatedly for classification in their breast cancer study. Adam et al. [5] and Qu et al. [6] used the "area under the Receiver Operating Characteristic (ROC) curve" criterion for feature selection and decision trees in conjunction with boosting techniques for classification in a prostate cancer study. Lilien et al. [7] analyzed the same data and the ovarian cancer data from Petricoin et. al. [2] with a probabilistic algorithm based on principal components and linear discriminant analyses.

In this paper we contrast the performance of several selected statistical and optimization-based techniques on the SELDI-TOF data from Adam et al. [5] and, in contrast to previous studies, assess their prediction significance via randomization techniques. These methods were selected for this study based on their simplicity and their widespread use and availability. While this is far from a comprehensive benchmark study, we make several points about the need for rigorous cross-validation and randomization.

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. Misclassification rates are biased (downward) when both the training and test sets are used for feature selection as opposed to when only the training set is used. While performing the cross-validation studies, care was taken so that the test set does not influence the choice of the peaks used in the classification.

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 [8] we have proposed candidate biomarkers in lung cancer using protein profiles obtained via MALDI-TOF (matrix assisted laser desorption and ionization-time

Table 1: Cross-validation classification accuracy (in percent) of various classification methods on the full four-class prostate cancer dataset using various numbers of peaks. Numbers are average observed accuracies over 100 runs with randomized 90/10 splits into training and test sets, respectively. The numbers in parentheses are the corresponding standard deviations.

	# of peaks used							
	10	15	20	25	30	35	50	70
Quadr. Discr.	74.7 (7.4)	74.7 (9.6)	74.1 (8.4)	74.7 (7.1)	78.2 (6.8)	77.8 (7.3)	78.7 (6.6)	76.8 (7.1)
Nonpar (Kernel)	76.7 (7.1)	77.4 (8.4)	77.7 (6.9)	78.6 (6.6)	80.0 (6.3)	79.9 (7.3)	78.1 (6.5)	76.1 (7.6)
kNN	73.4 (7.4)	76.4 (6.9)	76.9 (6.0)	76.6 (6.1)	75.8 (6.7)	77.2 (6.9)	73.9 (7.5)	69.8 (6.7)
Fisher Linear	72.4 (7.3)	77.3 (6.9)	80.8 (6.5)	80.1 (5.8)	81.8 (6.0)	84.6 (5.2)	85.5 (6.1)	84.3 (5.1)
Linear SVM	75.4 (6.4)	79.3 (7.4)	81.7 (7.2)	81.3 (5.7)	83.7 (6.8)	83.1 (6.6)	83.5 (6.1)	84.0 (6.2)

of flight) mass spectra (data provided by Duke University's department of Radiology) by employing similar data analysis strategies and classification methods. One of the five peaks found to be most useful in classifying lung cancer, and the only one up-regulated in cancer, was subsequently identified by Howard et. al [9] as stemming from Serum Amyloid A, an inflammatory marker, which we take to be an encouraging indication that this kind of analysis indeed has the potential to reveal relevant biomarkers.

Results and discussion

In what follows, unless otherwise indicated, the accuracy we report in the tables consists simply of the fraction of correctly classified samples, which is reasonable given that the sizes of the four class sizes are roughly balanced. Also, unless otherwise indicated, all error rates are computed as *average* error rates over 100 runs, that is, a cross-validation procedure of training on 90% of the data and testing on the remaining 10% was repeated 100 times and the errors averaged. Table 1 reports experiments on the original dataset with samples categorized into four groups: BPH, early (localized) cancer, late (metastasized) cancer and controls. Results for the first four classification methods (quadratic discrimination, nonparametric kernel method, kNN and Fisher's linear discriminator) reported were obtained with codes implemented in SAS. Results for the linear SVM were obtained with the package SvmFu [10]. For kNN we experimented with values of k between 3 and 7 and saw little overall sensitivity to the particular choice of k. We report results for $k = 6$. The linear SVM requires an *a-priori* choice of a tradeoff parameter C that balances misclassification and margin maximization. Instead of fine-tuning each SVM (which is rather computationally expensive, especially compared to the other four methods), we tried various discrete values ($\log C = -3, -2, \ldots, 1$) and observed that the best performance was always achieved with either $C = 1$ or $C = .1$. The results we report in the table correspond to the best of the these runs.

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 regard to the various phenotypes. Since there are four classes, a random classifier would be expected to achieve about 25% accuracy. We also note the rather high standard deviations (shown in parentheses), which indicate there was a wide range of observed classification accuracies over the 100 runs performed.

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 (the labels of the entire dataset were permuted at random) and averaged the performance of the linear 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 79.3% reported in Table 1 and is an indication that these results are not merely due to some spurious 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 are the subject of future research.

Table 2: Details of classification results obtained with Fisher's Linear Discriminator and 20 peaks on the full four-class problem. The overall average classification accuracy (100 runs) is 81%.

		Computational Prediction			
		BPH	Late Cancer	Early Cancer	Control
Clinical Diagnosis	BPH	745 (93.1%)	55 (6.9%)	0 (0%)	0 (0%)
	Late Cancer	156 (19.5%)	531 (66.3%)	91 (16.0%)	22 (1.6%)
	Early Cancer	99 (12.3%)	54 (6.8%)	616 (82.0%)	31 (1.8%)
	Control	92 (11.5%)	11 (1.4%)	5 (0.6%)	692 (86.5%)

Table 3: Average classification accuracy over 100 runs on data obtained by grouping all control and BPH samples into one class, and all cancer samples into another. Class sizes thus remain approximately balanced. Numbers in parentheses are standard deviations.

	# of peaks used (malignant vs. other)				
	5	8	10	12	15
Quadr. Disc.	84.1 (5.3)	85.1 (5.4)	85.0 (6.1)	86.1 (6.7)	86.0 (6.1)
Nonpar. (Kernel)	84.6 (5.2)	87.1 (5.3)	88.3 (5.8)	88.9 (6.1)	88.1 (6.0)
kNN	89.9 (4.6)	87.4 (5.6)	87.5 (5.7)	88.9 (5.2)	88.5 (4.6)
Fisher Linear	88.6 (5.9)	88.4 (5.6)	87.9 (4.9)	89.1 (5.4)	88.0 (5.0)
Linear SVM	89.5 (5.5)	91.0 (4.8)	91.9 (4.6)	91.7 (4.9)	91.9 (4.7)

For illustration purposes, we show detailed results obtained with Fisher's Linear Discriminator using 20 peaks on the full four-class problem in Table 2. We note that by far the largest source of misclassification comes from the late cancer group, indicating perhaps that it is a rather heterogeneous group in nature. 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.

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 (control or BPH) or cancerous. As seen in Table 3, this can be achieved with high accuracy (91.0%) with only 8 peaks using a linear SVM. Table 4 shows the average prediction accuracy achieved on other pairwise discriminations, indicating that the control versus BPH distinction can be made with 96% accuracy using just 5 peaks. Thus we obtain at least 87% accuracy for the two-stage process with a total of 13 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.

To assess the significance of these classification results we sampled from the empirical distribution of misclassification rates by randomly permuting the class labels. By performing 1000 randomization runs we can obtain estimates for the 95th percentile of that empirical error distribution. A generalization error estimate obtained by cross-validation on the true data set which is above the 95th percentile of the empirical error distribution can be interpreted as a confidence certificate. The following table contains the 95th percentile estimates and the means of the empirical misclassification error distribution for a subset of the methods and problems from above.

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.

We also want to mention briefly that when all 5 classification methods are trained using the entire dataset and 15 peaks, 74% of all samples are correctly classified by all methods simultaneously. We take this high level of concordance of the classification methods as a strong indication and additional evidence that a large majority of samples are indeed well separated in this low-dimen-

Table 4: Linear SVM classification average accuracy results for other pairwise distinctions using varying numbers of peaks.

	# of peaks used				
	5	8	10	12	15
BPH vs Control	96.4	96.2	96.6	96.4	97.4
BPH vs E. Cancer	91.8	94.6	93.6	94.7	95.4
BPH vs L. Cancer	89.1	88.1	88.9	89.7	91.7
Control vs E. Cancer	89.1	91.5	94.4	95.5	96.2
Control vs L. Cancer	88.0	88.7	88.5	90.4	90.0

Table 5: Statistics on classification accuracy for the linear SVM averaged over 1000 randomized datasets. 10 cross-validation runs using 15 peaks were performed on each dataset.

	max. acc.	median acc.	95th %ile
BPH vs Control	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
Control vs E. Cancer	66.9	50.0	59.7
Control vs L. Cancer	65.0	51.6	59.3

sional space, and that there is significant information content in this dataset which can be used to discriminate between the four classes.

Finally, we want to mention the top masses that repeatedly appear in the peak selection list of various classifiers: 9720.0, 9655.7, 5074.2, 3896.6, 3963.2, 7819.8, 7844.0, 6949.2, 8943.1, 4079.5. Some of these masses, e.g., 7819.8 and 9655.7, had also been used in previous studies (Adam et al., [5]) as being important discriminators. Note, however, that all masses are in the range of those of typical proteins (unlike those appearing in [2]). While the identification of the underlying proteins and our understanding of their biological significance is still outstanding, we believe that the results we provide here do indeed indicate that they are good candidates for biomarkers and that their identification can provide new insights of clinical relevance.

Methods

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), 84 patients with organ-confined prostate cancer (PCA), and 83 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. [5]. Each spectrum is an array of intensities of the signal at discretely sampled values of the mass-to-charge ratio of the ions.

In contrast to, e.g., Petricoin et. al. [2] our data is preprocessed by filtering out intensities that do not correspond to peaks in any samples since these will likely cloud the true information content of the spectrum. In order to use the intensities as indicators of relative abundance of the supposed peptide in the sample, baselines must be subtracted and the intensities normalized. Furthermore, mass measurement inaccuracies imply that peaks stemming from the same peptide may occur at different mass locations in different samples, so peaks must be aligned across samples. (See [8] for more details).

For the data in this paper, peak detection and alignment were performed with Ciphergen ProteinChip Software 3.0 with some modifications. All 652 spectra (326 samples in duplicate) were compiled and 779 peaks in the mass range from 2 to 40 kDa were selected by the ProteinChip software for analysis. This range contains the majority of the resolved protein/peptides [6]. Details of the steps involved in this pre-processing of the data are given in Adam et al. [5]. In this analysis, to avoid biasing the feature selection procedure and/or the cross-validation

results, we only used the data corresponding to the first mass spectrum for each sample.

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 a 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 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 appearing 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 more than 300 samples at our disposal we deemed this to be 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), for feature selection. Suppose that y_{ikj} is the observed intensity of the jth feature of the kth sample belonging to the ith group, that the number of groups is denoted by g, that n_i is the number of samples in the ith group, and that $\bar{y}_{ij} = \frac{1}{n_i}\sum_{k=1}^{n_i} y_{ikj}$, $\bar{y}_j = \frac{1}{\sum_{i=1}^{g} n_i}\sum_{i=1}^{g}\sum_{k=1}^{n_i} y_{ikj}$. Then the between group sum of squares for the jth feature is $B_j = \sum_{i=1}^{g}\left(\bar{y}_{ij} - \bar{y}_j\right)^2$ and the within group sum of squares is $W_j = \sum_{i=1}^{g}\sum_{k=1}^{n_i}\left(y_{ikj} - \bar{y}_{ij}\right)^2$. For every feature j = 1,...,p we compute B_j/W_j or, equivalently (for ordering purposes), the ANOVA F-statistic $F_j = \frac{B_j/\nu_1}{W_j/\nu_2}$, where ν_1 = g - 1 and $\nu_2 = \sum_{i=1}^{g} n_i - g$ are the degrees of freedom of B_j and W_j respectively. Then the reduced data set will be the data corresponding to the q largest F_j values.

Principal Component Analysis (PCA) is another popular and very general method for reducing the dimension of the data. Instead of working with the original variables, only a few selected "principal components," which are linear combinations of the original variables, are used for the analysis. Although the principal components have an advantage of explaining most of the variation in the original data, they may not be very useful in the present context where knowing the identity of the masses where the peaks occur is important. The elements of the eigenvectors that are used as principal components can, in principle, be used to select the original variables, but the heuristic and subjective nature of this approach makes this a less appealing approach for us in this context. (See Khattree and Naik [11] for an illustration of this approach and Lilien et al. [7] for an application to prostate cancer data.)

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. 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. [6]. Here we will adopt several classical statistical methods and support vector machines for our analysis. In the following we will only briefly mention these methods; details can be found in Khattree and Naik [11] and Hastie, Tibshirani, and Friedman [1].

The *quadratic discriminant rule* is a likelihood-based discriminant procedure in which multivariate normal probability density functions with unequal variance co-variance matrices are used. When the form of the probability density function is not known, the data can be used to estimate the densities. This is generally done using the kernel

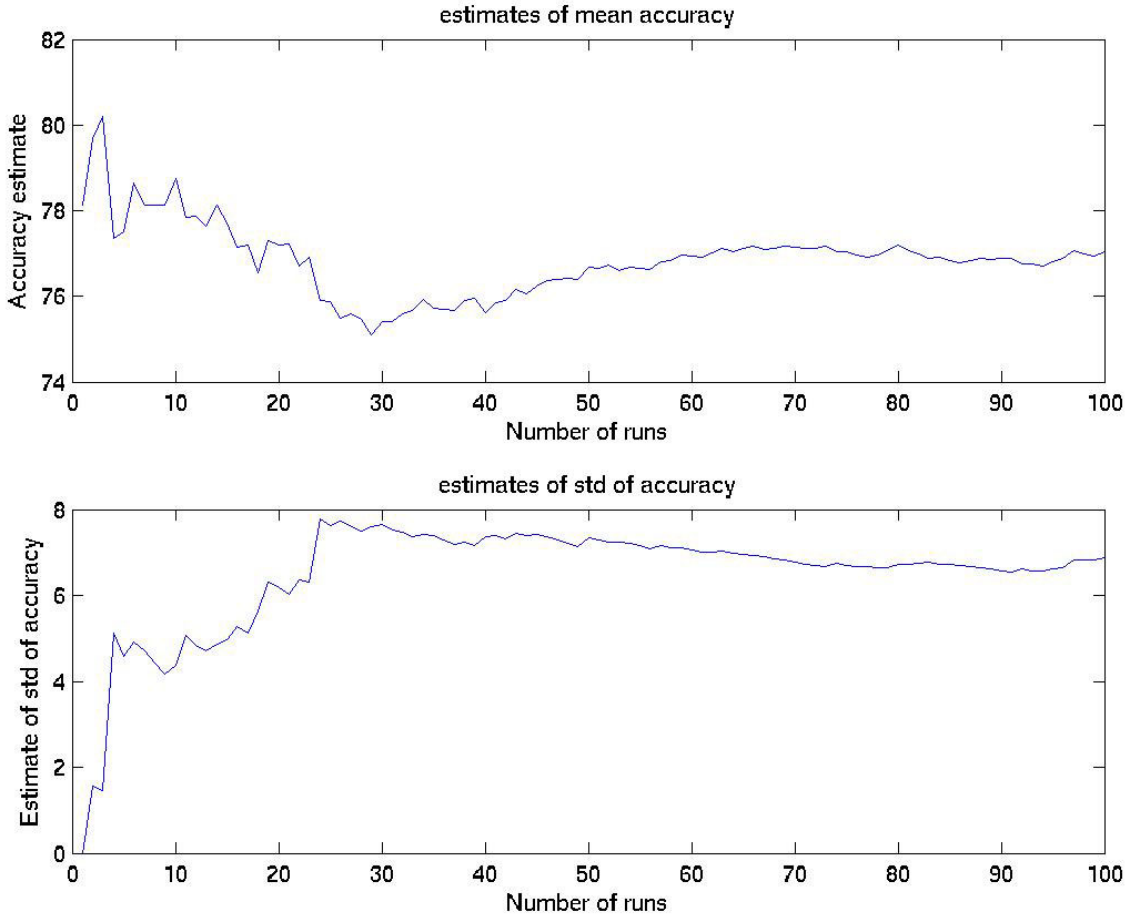

Figure 1
Accuracy and standard deviation estimates as a function of the number of cross-validation runs (shown, as an example, for the Fisher method with 15 peaks). Significant variability can be observed at the beginning, which motivates the need for a large number of runs in order to arrive at reasonable estimates.

method, and we will be referring to this procedure as *non-parametric (kernel) discrimination method*.

Fisher's canonical (linear) discriminant analysis provides a method where no probability density functions are used directly. 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 are likely to improve on the discrimination 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 k-nearest neighbor method (kNN) provides another approach that 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 known classified samples. Next, we find the samples corresponding to the k smallest values of the selected affinity measure, where k is pre-specified. The unknown sample is classified as belonging to the group to which the majority of the k closest samples

belong. 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., [12]). SVMs are large-margin classifiers, that is, they solve an optimization problem which finds a separating hyperplane that optimizes a weighted combination of the misclassification rate and the distance of the decision boundary to any sample vector. The reader is referred, e.g., to Cristianini and Shawe-Taylor [13] for details. In addition to linear versions of SVMs, they have been extended to nonlinear cases via kernels; however, since we did not see any significant performance improvement on this data using nonlinear kernels compared to the much simpler linear SVMs (data not shown), we constrained ourselves to reporting on experiments with linear SVMs. 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 [14] have found natural and theoretically satisfying extensions; however, we have opted for another scheme that is reasonable for small values of g, which, despite its simplicity, has produced quite satisfactory results in practice. We opted to adapt the pairwise approach that constructs all $g(g - 1)/2$ pairwise discriminators for g classes (groups). 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 that occurs in only a very small percentage of cases. We want to point out 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 clinical relevance.

Cross-validation

In order to assess the generalization power of each of the classification methods and to estimate their prediction capabilities for unknown samples, we used a standard cross-validation technique and split the data randomly and repeatedly into training and test sets. The training sets consisted of randomly chosen subsets containing 90% of each class (for a total of 294 per run); the remaining 10% of the samples from each class (a total of 32) 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 [8]) have shown that performing feature selection on the entire dataset often grossly underestimates the true generalization error.

Repeated cross-validation runs can be used to estimate the average classification accuracy as well as the standard deviation. However, obtaining reliable estimates (e.g., in order to compare different classification methods) is problematic and requires careful consideration. Figure 1 shows the variability of the estimates for mean classification accuracy as well as the standard deviation as a function of the number of cross-validation runs (using, as an example, Fisher's linear discriminator and 15 peaks). Even between runs 90 and 100 the estimates for mean error differ by as much as .4%, which indicates that any reported accuracies should really only be considered significant to the second digit. This is not really surprising: we performed a small simulation where 10,000 samples of size 100 were drawn from $N(13,7)$, and in 5% of cases the relative error of the sample mean (standard deviation) to the true mean (standard deviation) was greater than 10.6% (14.1 %), indicating that the mean and standard deviation estimates stemming from 100 runs are, in general, far from converged.

In order to keep computing times reasonable, we limited ourselves to reporting accuracy and standard deviation estimates over 100 runs, but we stress that more runs are required should more accurate estimates be desired.

Authors' contributions

MW experimented with SVMs and implemented and ran the cross-validation and randomization study; DNN worked with the statistical classification tools; AP, SK, and RRD processed the mass spectral data to extract peaks; and BLA, GLW, and OJS provided the mass spectral data; all authors contributed with the writing of the manuscript.

References

1. Hastie T, Tibshirani R, Friedman J: *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* Series in Statistics Springer; 2001.
2. Petricoin E III, Ardekani A, Hitt B, Levine P, Fusaro V, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC, Liotta LA: **Use of proteomic patterns in serum to identify ovarian cancer.** *The Lancet* 2002, **359:**572-577.
3. Sorace J, Zhan M: **data review and re-assessment of ovarian cancer serum proteomic profiling.** *BMC Bioinformatics* 2003, **4:**24.
4. Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW: **Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer.** *Clin Chem* 2002, **48:**1296-1304.
5. Adam BL, Qu Y, Davis JW, Ward MD, Clements MA, Cazares LH, Semmes OJ, Schellhammer PF, Yasui Y, Feng Z, Wright GL: **Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men.** *Cancer Res* 2002, **62:**3609-3614.
6. Qu Y, Adam BL, Yasui Y, Ward MD, Cazares LH, Schellhammer PF, Feng Z, Semmes OJ, Wright GL: **Boosted decision tree analysis of surface-enhanced laser desorption/lonlzatton mass spectral serum profiles discriminates prostate cancer from noncancer patients.** *Clin Chem* 2002, **48:**1835-1843.
7. Lilien R, Farid H, Donald B: **Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum.** *J Comp Biol* 2003, **10(6):**925-946.
8. Wagner M, Naik D, Pothen A: **Protocols for disease classification from mass spectrometry data.** *Proteomics* 2003, **3:**1692-1698.
9. Howard B, Wang M, Campa M, Corro C, Fitzgerald M, E Patz J: **Identification and validation of a potential lung cancer serum biomarker detected by matrix-assisted laser desorption/ion-**

ization-time of flight spectra analysis. *Proteomics* 2003, **3:**1720-1724.
10. Rifkin R: **SvmFu.** [http://five-percent-nation.mit.edu/SvmFu/].
11. Khattree R, Naik D: *Multivariate Data Reduction and Discrimination with SAS Software* SAS Institute and J Wiley and Sons; 2000.
12. Mangasarian OL: **Linear and nonlinear separation of patterns by linear programming.** *Oper Res* 1965, **13:**444-452.
13. Cristianini N, Shawe-Taylor J: *An Introduction to Support Vector Machines* Cambridge University Press, Cambridge, UK; 2000.
14. Lee Y, Lin Y, Wahba G: **Multicategory support vector machines.** *Comp Sci Stat* 2001, **33:**498-512.