Finding Diagnostic Biomarkers in Proteomic Spectra

Pallavi N. Pratapa, Edward F. Patz, Jr., Alexander J. Hartemink

Pacific Symposium on Biocomputing 11:279-290(2006)

FINDING DIAGNOSTIC BIOMARKERS IN PROTEOMIC SPECTRA

PALLAVI N. PRATAPA[1], EDWARD F. PATZ, JR.[2], ALEXANDER J. HARTEMINK[1]
[1]*Duke University, Dept. of Computer Science, Box 90129, Durham, NC 27708*
{pallavi,amink}@cs.duke.edu

[2]*Duke University Medical Center, Depts. of Radiology & Pharmacology and Cancer Biology, Box 3808, Durham, NC 27710*
patz0002@mc.duke.edu

In seeking to find diagnostic biomarkers in proteomic spectra, two significant problems arise. First, not only is there noise in the measured intensity at each m/z value, but there is also noise in the measured m/z value itself. Second, the potential for overfitting is severe: it is easy to find features in the spectra that accurately discriminate disease states but have no biological meaning. We address these problems by developing and testing a series of steps for pre-processing proteomic spectra and extracting putatively meaningful features before presentation to feature selection and classification algorithms. These steps include an HMM-based latent spectrum extraction algorithm for fusing the information from multiple replicate spectra obtained from a single tissue sample, a simple algorithm for baseline correction based on a segmented convex hull, a peak identification and quantification algorithm, and a peak registration algorithm to align peaks from multiple tissue samples into common peak registers. We apply these steps to MALDI spectral data collected from normal and tumor lung tissue samples, and then compare the performance of feature selection with FDR followed by classification with an SVM, versus joint feature selection and classification with Bayesian sparse multinomial logistic regression (SMLR). The SMLR approach outperformed FDR+SVM, but both were effective in achieving good diagnostic accuracy with a small number of features. Some of the selected features have previously been investigated as clinical markers for lung cancer diagnosis; some of the remaining features are excellent candidates for further research.

1 Introduction and motivation

A diagnosis of cancer is often first suggested by radiological imaging. Unfortunately, imaging findings are not always specific so further evaluation with invasive procedures is typically required to establish a diagnosis. Many researchers have put great effort into developing alternative strategies for more effectively diagnosing cancer non-invasively, particularly through the identification of diagnostic biomarkers. While some groups have focused on genomics, others have pursued proteomics, hoping that protein expression profiles will lead to biomarkers that more accurately reflect disease phenotypes. Given the limitations of traditional methods involving 2D-GE, alternative proteomic platforms have been pursued. Over the last several years investigators have begun to explore the use of a variety of protein separation techniques followed by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry (MALDI-TOF MS; henceforth just 'MALDI'). Although MALDI has traditionally been used for protein identification, several recent studies have suggested that direct analysis of MALDI data can provide diagnostic value.[1,2]

The data from MALDI is a list of mass-charge ratios ('m/z values') and cor-

responding measured intensities. If we plot the measured intensities as a function of m/z, we call the resultant curve a 'spectrum'. Peaks in the spectrum correspond to proteins in the tissue sample. Under ideal conditions, samples with similar protein composition would have peaks with identical intensities at identical m/z values. However, due both to variation in lysate preparation and limitations inherent in the measurement technology, not only is there noise in the measured intensity at each m/z value, but there is also noise in the measured m/z value itself. This makes it difficult to directly compare spectra between groups of patients and thus to identify specific protein expression patterns from complex biological samples. Because the spectral data we collect possess a hierarchical structure (multiple replicate spectra per sample, multiple samples per class), we develop a hierarchical strategy for solving this problem. We use a latent spectrum extraction algorithm to fuse information from multiple replicate spectra obtained from a single tissue sample, and then a peak registration algorithm to align peaks from multiple tissue samples into common 'peak registers'. These two algorithms are embedded in a longer data analysis pipeline (described below), designed to identify a small number of features as putatively meaningful diagnostic biomarkers. While previous methods have been developed for particular steps in this pipeline, and while very recent reviews have admirably and effectively summarized previously published methods,[3,4] our experience in implementing the entire pipeline has enabled us to test and compare both existing and novel methods at each step in the analysis, and in the context of the full hierarchical pipeline. Here, for reasons of limited space, we report the final choices that were made at each step.

2 Analytical methods

2.1 Overview

Our data analysis pipeline is hierarchically organized and consists of two levels of pre-processing followed by a third level of feature selection and classification. The first pre-processing level identifies and quantifies putatively meaningful peaks in each tissue sample from multiple replicate spectra. The second pre-processing level yields a matrix of comparable features across all the tissue samples. The steps in each of these levels of analysis are depicted in Fig. 1. In what follows, we discuss each of these steps in turn, devoting more attention to the more interesting steps.

2.2 Latent spectrum extraction

Ideally, the multiple replicate spectra collected from each tissue sample would be identical, but unfortunately, both X- and Y-axis measurements are noisy. As a model, we postulate that the measured spectra are noisy versions of some true 'latent spectrum'. We imagine that variability in the Y-axis arises from a combination of multiplicative and additive errors: a global scaling factor for each replicate, a local scaling

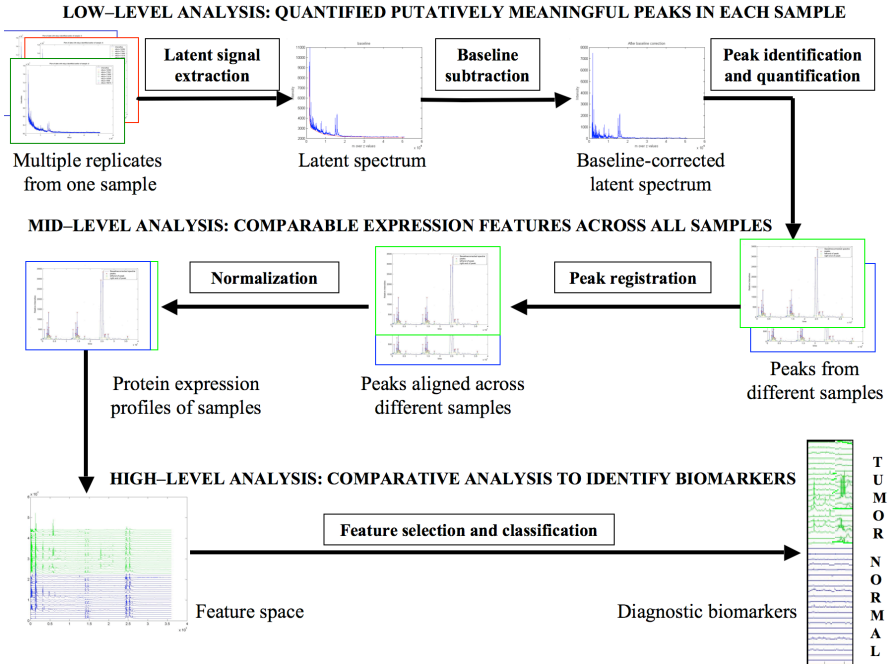

Figure 1. Overview of our hierarchical data analysis pipeline.

factor which varies smoothly within each replicate, and a local additive noise term. The global scaling factor can account for things like variation in the concentration of ions present at the laser location corresponding to each replicate spectrum. Likewise, the local scaling factor can account for signal suppression factors which influence the number of protein ions passing through the detector at each point in time. We further imagine that variability in the X-axis arises from non-uniform subsampling of the latent spectrum, which is different for each of the replicates. A latent spectrum extraction algorithm can then be used to find the latent spectrum that maximizes the likelihood of the measured replicate spectra. For this purpose, we use the recently-proposed continuous profile model (CPM),[5] implementing a learning algorithm with a few computational enhancements for our setting. Below, we formalize the model, provide brief details of the learning algorithm, and illustrate its operation with an example; readers interested in further details are encouraged to read the original paper.[5]

2.2.1 Model description

To ease comparison with the original paper, we adopt nearly identical notation, although we do correct a few errors. Assume we have K replicate spectra, indexed by $k \in \{1, 2, \ldots, K\}$. Let $X^k = [x_1^k, x_2^k, \ldots, x_{N^k}^k]$ denote the measured intensity val-

ues in the k-th replicate spectrum at the m/z values indexed by $i \in \{1, 2, \ldots, N^k\}$. Let $Z = [z_1, z_2, \ldots, z_M]$ denote the (unmeasured) intensity values in the latent spectrum at the m/z values indexed by $\tau \in \{1, 2, \ldots, M\}$. According to the model, each replicate spectrum is a non-uniformly subsampled version of the latent spectrum, to which global and local scaling factors have been applied and noise has been added. Hence, we have

$$x_i^k = z_{\tau_i^k} \phi_i^k u^k + \epsilon \tag{1}$$

where τ_i^k is the hidden time state (the value of τ in the latent spectrum that corresponds with i in replicate k), ϕ_i^k is the hidden local scale state, u^k is the global scaling factor for replicate k, and ϵ is drawn from a central normal distribution with variance σ^2, assumed to be the same for all replicates.

As previously described,[5] this is essentially a hidden Markov model (HMM), where the output is conditioned on the latent spectrum. Let π^k denote the hidden state sequence for the k-th replicate. Each state in this sequence consists of a time state and a local scale state: $\pi_i^k = \langle \tau_i^k, \phi_i^k \rangle$. The time states are selected from the m/z values of the latent spectrum ($\tau_i^k \in \{1, 2, \ldots, M\}$) and the local scale states are selected from an ordered set of P scale values ($\phi_i^k \in \{\phi^1, \phi^2, \ldots, \phi^P\}$). The HMM is specified by the following emission and transition probabilities:

$$\textbf{Emission:} \quad e_{\pi_i^k}(x_i^k | Z) = P(x_i^k | \pi_i^k, Z, u^k, \sigma^2) = \mathcal{N}(x_i^k; z_{\tau_i^k} \phi_i^k u^k, \sigma^2) \tag{2}$$

$$\textbf{Transition:} \quad T^k_{\pi_{i-1}, \pi_i} = P^k(\pi_i | \pi_{i-1}) = P^k(\tau_i | \tau_{i-1}) P(\phi_i | \phi_{i-1}) \tag{3}$$

Regarding the transition probabilities between time states, we impose the constraint that the time state must always advance at least one step and no more than J steps from the current state; transitioning $v \in \{1, \ldots, J\}$ steps is multinomial with probability d_v^k. Regarding the local scale states, we impose the constraint that the scale state must remain the same or change to the neighboring scale state above or below; transitioning $v \in \{-1, 0, 1\}$ steps is multinomial with probability $s_{|v|}$.

2.2.2 Learning the latent spectrum using expectation-maximization

Given the replicate spectra from a single tissue sample, we can train the model to learn the latent spectrum using Baum-Welch (EM). The M-step update rules are derived by solving for the values of the parameters Z, σ^2, and u^k that maximize the expected log likelihood (we ignore the smoothing prior), yielding the following:

$$z_j = \frac{\sum_{k=1}^K \sum_{\{s|\tau_s=j\}} \sum_{i=1}^N (\gamma_s^k(i) x_i^k \phi_s u^k)}{\sum_{k=1}^K \sum_{\{s|\tau_s=j\}} \sum_{i=1}^N (\gamma_s^k(i) (\phi_s u^k)^2)} \tag{4}$$

$$\sigma^2 = \frac{\sum_{k=1}^K \sum_{s=1}^S \sum_{i=1}^N \gamma_s^k(i) (x_i^k - z_{\tau_s} \phi_s u^k)^2}{KN} \tag{5}$$

$$u^k = \frac{\sum_{s=1}^S z_{\tau_s} \phi_s \sum_{i=1}^N \gamma_s^k(i) x_i^k}{\sum_{s=1}^S (z_{\tau_s} \phi_s)^2 \sum_{i=1}^N \gamma_s^k(i)} \tag{6}$$

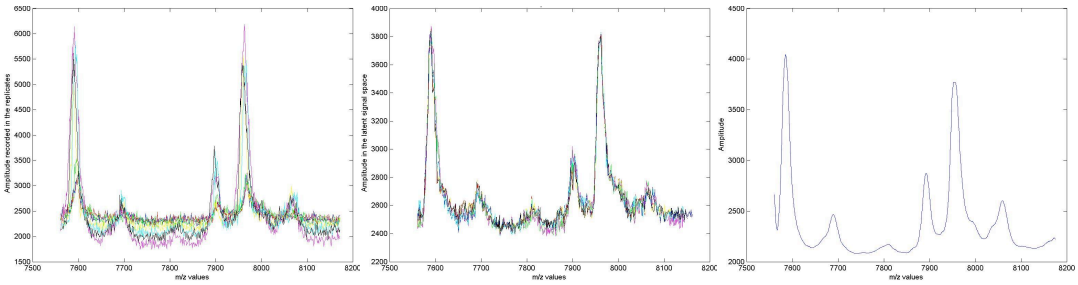

Figure 2. Latent spectrum extraction. (a) A subportion of length N=500 of seven unaligned replicate spectra from a single tissue sample; note variation in both the X- and Y-axes. (b) The replicate spectra after Viterbi alignment to the latent spectrum. (c) The latent spectrum extracted from the replicate spectra.

where $s \in \{1, 2, \ldots, S\}$ indexes the total number of possible states and $\gamma_s^k(i) = P(\pi_i = s|x_i^k)$ is computed during the forward-backward algorithm in the E-step. Similarly, the update rules for the various multinomial transition probabilities are shown below:

$$d_v^k = \frac{\sum_{s=1}^{S} \sum_{\{s'|\tau_s - \tau_{s'} = v\}} \sum_{i=2}^{N} \xi_{s,s'}^k(i)}{\sum_{s=1}^{S} \sum_{j=1}^{J} \sum_{\{s'|\tau_s - \tau_{s'} = j\}} \sum_{i=2}^{N} \xi_{s,s'}^k(i)} \tag{7}$$

$$s_v = \frac{\sum_{k=1}^{K} \sum_{s=1}^{S} \sum_{\{s' \in H(s,v)\}} \sum_{i=2}^{N} \xi_{s,s'}^k(i)}{\sum_{k=1}^{K} \sum_{s=1}^{S} \sum_{\{s' \in H(s,0), H(s,1)\}} \sum_{i=2}^{N} \xi_{s,s'}^k(i)} \tag{8}$$

where $H(s, v)$ indicates the set of states that are exactly v scale values away from s and $\xi_{s,s'}^k(i) = P(\pi_{i-1} = s, \pi_i = s' \mid x_i^k)$ is again computed during the E-step.

2.2.3 Illustration of latent spectrum extraction

We illustrate the results of applying the latent spectrum extraction algorithm to fuse information from multiple replicate spectra from a single sample. We consider a section of N=500 points of K=7 unaligned replicate spectra of a sample as shown in Fig. 2(a). For training the model, we set $M = 2.01 \times N$. The latent spectrum Z is initialized to be the median of the intensity values in the replicates, with Gaussian noise added; the standard deviation of the Gaussian noise is initialized to 30% of the difference between the minimum and maximum value in Z. The result is supersampled by a factor of two (repeating every value twice consecutively) with additional small values padding the beginning and end to achieve a total length of M. For each replicate, the global scale value u^k is initialized to unity. The largest time state transition, J, is set to 3. An ordered set of P=21 local scale values, $\{0.80, 0.82, 0.84, \ldots, 1.20\}$, is used. The multinomials defining the time state and scale state transition probabilities are initialized to uniform. The transition probabilities from the beginning state to π_1^k are set uniformly.

After training the model, we use the Viterbi algorithm to find the most likely path

through the hidden states for each replicate. We then align each of the replicates to the latent profile, the output of which is shown in Fig. 2(b). As the figure illustrates, the replicates are now all well-aligned on both the X-axis and the Y-axis. Fig. 2(c) shows the latent spectrum, which offers a denoised summary representation of the information contained in the unaligned replicate spectra.

2.3 Baseline subtraction

To compensate for the gradually decreasing baseline of a complete latent spectrum, we find a monotone local minimum curve that lower-bounds it by computing the convex hull of the latent spectrum. We then subtract this from the latent spectrum to get the baseline-corrected latent spectrum, or simply latent spectrum henceforth. In addition to its extreme simplicity, an advantage of this method over previously proposed methods is that the resultant latent spectrum is everywhere non-negative.

2.4 Peak identification and quantification

Identifying peaks in the latent spectrum of a tissue sample is imperative because peaks represent the isolatable proteins or peptides that may be relevant in discriminating between the two classes of samples. Focusing our attention on the peaks rather than the entire spectrum eliminates from consideration potential features that are expected to have no biological meaning.

To identify the important peaks in a latent spectrum, a simple approach would be to identify all locally maximal points in the spectrum with height above a certain signal-to-noise threshold, T. However, a complication arises because some local maxima satisfying this criterion are simply noisy bumps on the side of a larger peak. For this reason, and because later we will quantify peaks by their area, we also need to determine an interval describing the support (on the m/z axis) of each putative peak. To accomplish this, we develop a peak picking algorithm that begins by sorting all local maxima in the latent spectrum into a priority queue by their signal-to-noise ratio, truncating the queue when the ratio drops below T. The extent of the support interval is determined for the peak at the head of the queue by moving left and right from the peak until a moving average of the gradient changes sign (the moving average prevents us from stopping at noisy bumps on the side of a larger peak). Elements of the queue falling within this interval are removed from the queue. Processing the entire queue gives us a list of putatively important peaks, along with a height and width for each.

The height of a peak is not good for quantifying its relative importance in the spectrum. For fixed laser energy, recorded intensity values are generally higher for lower m/z values; the peaks at lower m/z values are also narrower than those at higher m/z values. Hence, the area enclosed by a peak within its support interval has been suggested as perhaps a more useful measure for comparing peaks over a wide range of m/z values. This area is the measure we use for peak quantification.

2.5 Peak registration to align peaks across all samples

Just as m/z values for a peak can vary within replicates from the same tissue sample, they can also do so across samples. However, now it is not the case that we expect a single underlying latent spectrum to explain the spectra from different samples because the samples themselves may be biologically heterogeneous. To address this problem and enable peaks from different samples to be compared on an equal footing, we develop a simple peak registration algorithm, developed independently of but similar in flavor to one recently published.[6] Given a list of peaks from the latent spectra of all the samples, and an estimate of the mass resolution error of the MALDI instrument, we assign peaks in different latent spectra into the same 'peak register' if their m/z values are within the mass resolution of the instrument. Since the instrument's resolution is proportional to m/z, we first log-transform the m/z values and then perform hierarchical clustering in log-m/z space using complete linkage and a Euclidean distance metric. We can determine the number of registers by cutting the dendrogram at a depth given by the log-transformed mass resolution of the instrument. The m/z value that we associate with each peak register is the mean of the m/z values of the peaks that belong to the register.

2.6 Normalization

The areas of peaks belonging to one m/z register may have a high coefficient of variation across different samples of similar protein composition. The areas of the peaks for each sample need to be normalized to make a fair comparison of protein expression levels across different samples. We normalize by a global factor, computed so as to equalize either the mean peak area or the median peak area of all samples. In each case, the normalized peak areas are finally log-transformed so as to not overemphasize obvious peaks in relation to less obvious ones. This step could also be performed before the previous step because the previous step takes no account of the peak areas. More sophisticated strategies might merge these two steps.

2.7 Feature selection and classification

Once we have identified a matrix of comparable features across all the tissue samples, we can consider strategies for sparse feature selection and classification. If we rank each feature based on its Fisher discriminant ratio (FDR), one strategy is to start with the top ranking feature and sequentially add features until there is no further improvement in the leave-one-out cross-validation ('LOOCV') classification accuracy of a linear SVM. This strategy has the benefit of producing a classifier with a small number of features through the sequential combination of two common methods: FDR for feature selection and SVM for classification.

However, we may be able to do better by using a single method that jointly addresses the tasks of feature selection and classification, especially in a proteomic context where features so severely outnumber observations. Bayesian algorithms

for learning sparse classifiers have recently been developed to learn simultaneously both a small subset of features relevant to classification and an optimal classifier.[7,8,9] Sparsity-promoting priors are used to regularize the feature weight vector, ensuring that weights are either significantly large or exactly zero, automatically removing irrelevant features from consideration. We use sparse multinomial logistic regression (SMLR).[9] The sparsity of the feature weight vector is controlled by the regularization parameter λ. Too high a value of λ will result in relevant features not being selected, thereby giving rise to more errors during training and cross-validation. On the other hand, too small a λ will cause more features to be selected than should be, resulting in over-fitting during training and more errors during cross-validation. Consequently, we can choose λ using LOOCV.

3 Results

3.1 Experimental procedure for collecting MALDI spectral data

Resections of lung tissue were obtained from 34 patients diagnosed with non-small cell lung cancer. In each case, normal and tumor lung tissue samples from the same patient were collected, yielding 68 total samples. Tissue samples were washed to remove blood contamination and then placed into a micro-centrifuge tube containing a protein extraction reagent and electrically homogenized. Cellular debris was removed by centrifugation and cell lysates were prepared. One microliter of lysate was deposited on the MALDI matrix and allowed to dry under ambient conditions. Spectra were acquired on a Voyager DE Biospectrometry Workstation using a nitrogen laser (337 nm). Multiple spectra were obtained from each tissue sample by focusing the laser on different sub-positions on the deposited lysate. By visually examining each spectrum for sufficient signal in terms of the number of peaks, ten replicate spectra were chosen for each tissue sample, for a total of 680 spectra.

3.2 Extraction of the latent spectrum from the replicates of a sample

Each replicate spectrum of each tissue sample contains N=27715 data points, with m/z values ranging from 1500 to 44000, approximately. We split the X-axis of the replicates into four sections to increase the speed of processing the data: we can run the latent spectrum extraction algorithm in parallel over the four sections of the replicate spectra. The spectra are split such that peaks exist in each section, but the tail end of each section contains only noise; this enables the sections to be recombined without loss of information about the peaks.

The various parameters specifying the latent spectrum extraction model and initializing the learning algorithm were chosen exactly as in the example of Sec. 2.2.3, except that for computational reasons, only three local scale values $\{0.8, 1, 1.2\}$ were used, representing local down-scaling, no scaling, and up-scaling of the latent spectrum, respectively. The model was trained using EM to learn the latent spectrum for

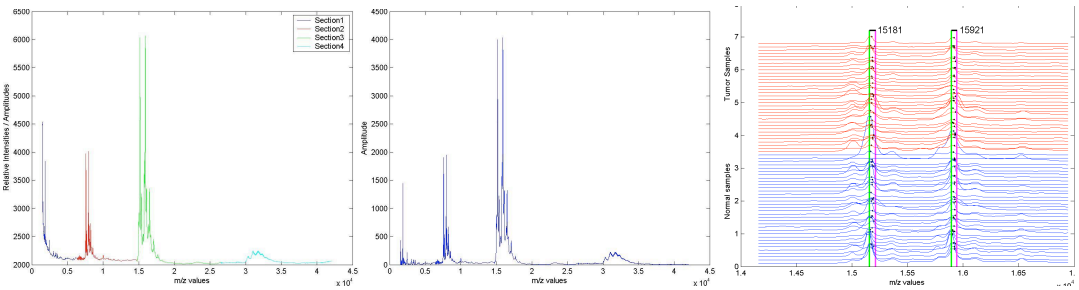

Figure 3. (a) The recombined latent spectrum of a sample obtained from ten replicate spectra. The four sections are shown in four different colors. (b) The baseline-corrected latent spectrum. (c) An example of peak registration, showing how peaks from multiple latent spectra are assigned to common peak registers.

each of the four sections. The four sections of the latent spectra were then combined by attaching them end to end and overlapping the last six points of one section with the first six points of the following section. As an example, the final recombined latent spectrum from the ten replicates of one tissue sample is shown in Fig. 3(a). The baseline is found using a segmented convex hull and subtracted from the latent spectrum to obtain the baseline-corrected latent spectrum, as shown in Fig. 3(b).

3.3 Peak registration

Cutting the complete-linkage hierarchical clustering dendrogram at a height equal to the log-transformed mass resolution error of the instrument produces 380 peak registers. In Fig. 3(c), the results of peak registration on two peak registers at m/z values 15181 and 15920 are shown, along with the boundary of each register; the peaks at these two m/z values are known to be present in all samples.

3.4 Normalization

We compared the two normalization strategies based on how much the standard deviations of the peak areas for the normal and tumor samples, taken as a class, were reduced by normalization in each case. Before normalization, the standard deviations of the peak areas for normal and tumor samples were 156 and 137, respectively. After normalization using mean peak area, these values became 196 and 110, respectively. After normalization using median peak area, these values became 128 and 72, respectively. Better performance using the median is reasonable since the mean is more susceptible to outliers; as a result of this consideration, and the corroborating numerical results, we proceed with normalization based on median peak areas. As mentioned above, the normalized peak areas were finally log-transformed so as to not overemphasize obvious peaks in relation to less obvious ones.

To summarize the two levels of pre-processing, whereas our initial data consisted of 680 replicate spectra of length N=27715, this has now been transformed

Table 1. LOOCV classification accuracy of SVM and SMLR based on an optimal set of selected features.

FDR+SVM		SMLR	
Accuracy	# features	Accuracy	# features
82% (9 errors)	7	92% (4 errors)	4

Table 2. m/z values of features selected by FDR and SMLR (ordered by rank).

Features selected by FDR	Features selected by SMLR
12386, 17932, 6206, 10884, 11215, 9781, 9109	12386, 5147, 16140, 15378

into a matrix with one row for each of the 68 tissue samples, one column for each of the 380 identified peak registers, and entries representing log-transformed normalized peak areas.

3.5 Feature selection and classification

We assess the performance of both feature selection and classification strategies discussed previously. For the first strategy, we sequentially add features based on their FDR ranking and continue in this manner until we no longer improve our generalization performance with a linear SVM classifier, in terms of LOOCV error. For the second strategy, we vary the value of λ over a range of values and select one that optimizes our generalization performance based on sparse multinomial logistic regression.

Table 1 summarizes the LOOCV classification accuracy of the two different classifiers with the subset of features selected to minimize LOOCV error. The performance of SMLR can be seen to be noticeably better than SVM on this particular data. Table 2 lists the m/z values of the features selected by both methods, ordered by rank. The top feature in each case is the same, indicating that it provides good discriminatory information between the two classes when used alone.

The plot on the left of Fig. 4 shows a grayscale representation of the features selected by FDR+SVM. The X-axis indexes the selected features and the Y-axis indexes the 68 samples, sorted by class. Each bar represents the log-transformed normalized peak area of the peak register in each sample. The features have clearly differential patterns of expression across the two different classes. The plot on the right shows the margin of each sample under SVM; blue open circles and red filled triangles represent normal and tumor classes respectively. Similar plots are shown for SMLR in Fig. 5. As can be seen from the plot on the left, SMLR selects features with non-redundant expression patterns, in contrast to FDR which picks features with redundant expression patterns; this is as expected. The plot on the right shows the probability of each sample being normal under SMLR.

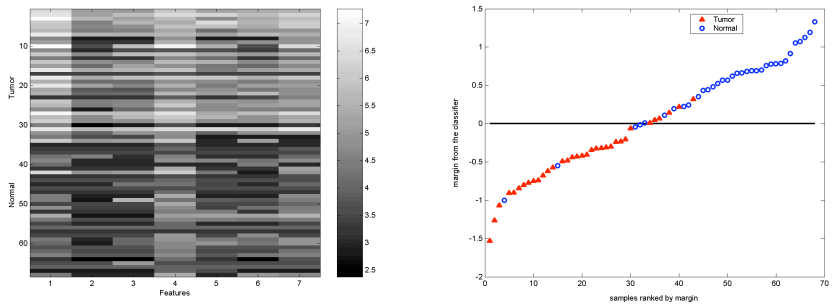

Figure 4. Features selected by FDR and classification by SVM using the selected features. The plot on the left depicts the differential patterns of the selected features across the samples belonging to the two different classes. The plot on the right shows the margin of each sample under SVM.

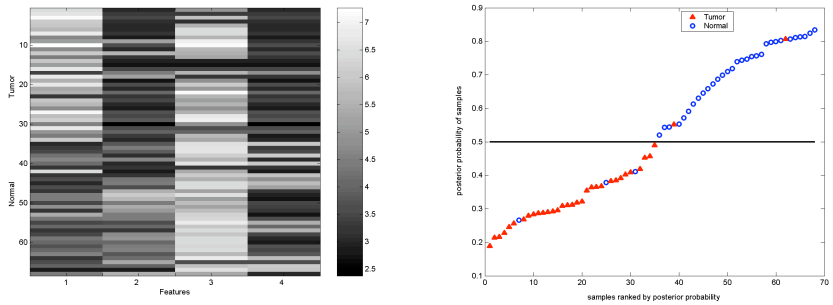

Figure 5. Feature selection and classification by SMLR. The plot on the left depicts the differential patterns of the selected features across the samples belonging to the two different classes. The plot on the right shows the probability of each sample being normal under SMLR.

4 Discussion

Even assuming that the data analysis pipeline presented here is successful in finding putatively meaningful discriminatory features, as it was in this case, isolation and identification of the proteins or peptides corresponding to these features must still be undertaken. In the case of this particular dataset, the two proteins found at m/z values 12386 and 17932 have already been identified by immunohistochemical analysis as macrophage migration inhibitory factor and cyclophilin A, respectively;[10] the prognostic value of these markers is currently under investigation. Proteins corresponding to certain other features are also being identified.

A number of improvements could be made upon the methods presented herein. Our methods ignore that ions can be multiply-charged, leading to the presence of

'harmonic' peaks in the spectra. For example, peaks at m/z values 12386 and 6206 were both found by FDR to be overexpressed in the tumor tissue samples, but these may be singly- and doubly-charged variants of the same protein. In addition, our methods also ignore isotope variants of a molecule[11] which, although chemically identical, can result in the presence of 'sister' peaks near the peak of the predominant isotope. Because SMLR selects non-redundant features, it would seem less prone to this sort of problem, but incorporation of this information into the peak identification and quantification step can only help.

Our peak registration algorithm aligns peaks based on only on their m/z locations, and ignores information about the heights, widths, or even shapes of peaks. To incorporate this information, alignment algorithms similar to multiple sequence alignment or dynamic time warping (DTW) for multiple alignment of speech signals could be used.

Finally, in the hierarchical strategy we proposed, the output of one step is piped as the input to the next step sequentially. Instead, a more unified model that combines different steps of analysis could provide a framework to share information across different steps, possibly leading to better results than is possible with our sequential approach.

[Larger versions of all figures are available from http://www.cs.duke.edu/∼amink/]

References

1. G. L. J. Wright. *Expert Rev Mol Diagn*, 2:549–563, 2002.
2. S. A. Schwartz, M. L. Reyzer, and R. M. Caprioli. *J Mass Spectrom*, 38:699–708, 2003.
3. S. Hyunjin and M. K. Markey. *J Biomedical Informatics*, 38, 2005.
4. J. Listgarten and A. Emili. *Molecular and Cellular Proteomics*, 4:419–434, 2005.
5. J. Listgarten, R. M. Neal, S. T. Roweis, and A. Emili. In *Advances in Neural Information Processing Systems*, volume 17, Cambridge, MA, 2005. MIT Press.
6. R. Tibshirani, T. Hastiey, N. Balasubramanian, S. Soltys, G. Shi, A. Koong, and Q. T. Le. *Bioinformatics*, 2004.
7. M. Tipping. *J Machine Learning Research*, 1:211–244, 2001.
8. M. Figueiredo and A. Jain. In *Computer Vision and Pattern Recognition*, 2001.
9. B. Krishnapuram, M. Figueiredo, A. J. Hartemink, and L. Carin. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:957–968, 2005.
10. M. J. Campa, M. Z. Wang, B. Howard, M. C. Fitzgerald, and E. F. J. Patz. *Cancer Res*, 63:1652–1656, 2003.
11. K. R. Coombes, J. M. Koomen, J. S. Baggerly, K. A.and Morris, and R. Kobayashi. Technical report, Department of Biostatistics and Applied Mathematics, UT M.D. Anderson Cancer Center, 2004.