Data and text mining

Analysis of mass spectral serum profiles for biomarker selection

Habtom W. Ressom1,#, Rency S. Varghese1, Mohamed Abdel-Hamid2, Sohair Abdel-Latif Eissa3, Daniel Saha1, Lenka Goldman1, Emanuel F. Petricoin4, Thomas P. Conrads5, Timothy D. Veenstra5, Christopher A. Loffredo1 and Radoslav Goldman1

1Lombardi Comprehensive Cancer Center, Georgetown University, Washington, DC, USA, 2Viral Hepatitis Research Laboratory, NHTMRI, Cairo, Egypt, 3National Cancer Institute, Cairo, Egypt, 4Clinical Proteomics Program, NCI/FDA, Center for Biologics Evaluation, FDA, USA and 5SAIC-Frederick and Biomedical Proteomics Program, NCI, Frederick, MD, USA

Received on May 24, 2005; revised on August 30, 2005; accepted on September 8, 2005
Advance Access publication September 13, 2005

ABSTRACT
Motivation: Mass spectrometric profiles of peptides and proteins obtained by current technologies are characterized by complex spectra, high dimensionality and substantial noise. These characteristics generate challenges in the discovery of proteins and protein-profiles that distinguish disease states, e.g. cancer patients from healthy individuals. We present low-level methods for the processing of mass spectral data and a machine learning method that combines support vector machines, with particle swarm optimization for biomarker selection.

Results: The proposed method identified mass points that achieved high prediction accuracy in distinguishing liver cancer patients from healthy individuals in SELDI-QqTOF profiles of serum.

Availability: MATLAB scripts to implement the methods described in this paper are available from the HWR’s lab website http://lombardi.georgetown.edu/labpage
Contact: hwr@georgetown.edu

1 INTRODUCTION
Mass spectrometric profiling of serum was optimized for high-throughput comparison of complex samples that allows discovery of biomarkers of diseases such as cancer (Petricoin et al., 2002a). Independent analysis of the results pointed out the importance of avoiding bias and the need for independent validation of results (Baggerly et al., 2004; Diamandis, 2004; Ransohoff, 2005). The improved study design and technology used in second-generation studies continue to indentify biomarker-candidates for a variety of cancers, including hepatocellular carcinoma (Zhang et al., 2004; Conrads et al., 2004; Paradis, 2005). This paper adds signal processing and biomarker selection methods to a growing number of improved tools for mass spectrometric identification of biomarkers in serum.

Data preprocessing, such as smoothing, baseline correction, normalization, peak detection and peak alignment, improve the performance of mass spectrometric data analysis methods for biomarker discovery (Sauve and Speed, 2004; Malyarenko et al., 2005). The reason for this includes the substantial amount of noise and systematic variations between spectra caused by sample degradation over time, ionization suppression and other parameters reviewed previously (Ransohoff, 2005; Semmes, 2005). Sorace and Zhan (2003) have reported the potential for non-biologic experimental bias in their assessment of ovarian cancer serum surface-enhanced laser desorption/ionization time-of-flight (SELDI-TOF) profiling due to the presence of very low mass classifiers, which the authors concluded could not possibly come from biology. This contention was disputed since many low molecular weight molecules detected by mass spectrometry could be metabolites or entities, such as lysophosphatidic acid, a potential biomarker for ovarian cancer detection, with a mass of 430 Da (Petricoin et al., 2004, www.biomedcentral.com/1471-2105/4/24/comments).

Mass spectra represent a complex signal consisting of electronic noise, chemical noise due to contaminants and matrix, and protein and metabolic signatures (Petricoin et al., 2002b). They also have a varying baseline caused, besides others, by matrix-associated chemical noise or by ion overload. The latter refers to the high excess of ions derived from the matrix that can overload the detector (Malyarenko et al., 2005). This elevates the baseline from its ideal zero horizontal line.

Previous quality-control experiments have suggested several measurement properties of current mass spectrometry technologies that must be accounted for in the analysis (Yasui et al., 2003). These properties include high dimensionality of the spectra, high coefficients of variation and mass shift (measurement error) Thus, it is important to apply low-level analyses that enable the recognition of spectral quality prior to using the spectra for biomarker discovery and disease classification. The low-level corrections are typically available in every software for the operation of a mass spectrometer. The use of spectral comparisons for biomarker identification requires, however, optimization of these methods and a completely transparent data manipulation. Several groups proposed recently improved tools for signal processing for biomarker discovery as summarized briefly below.

By smoothing the raw spectra, we can reduce the effect of some mass-per-charge (m/z) values that appear as peaks but these may not be easy or are very hard to be verified by independent experiments.

#To whom correspondence should be addressed.
Many smoothing algorithms are available to denoise raw signals, including the well-known Savitzky-Golay filter that removes additive white noise (Pusztai et al., 2004) and wavelets (Coombes et al., 2004).

Baseline correction is important for minimization of background noise; drifting baseline introduces serious distortion of ion intensities without adequate correction. Several methods have been proposed for baseline subtraction. For example, Fung and Enderwick (2002) employed a varying-width segmented convex hull algorithm to subtract the baseline. Baggerly et al. (2003) fitted a local median or a local mean in a fixed window on the time scale. They also considered subtracting a ‘semimonotonic’ baseline. Coombes et al. (2004) estimated the baseline by fitting a monotone local minimum curve to smoothed spectra.

Normalization reduces variation in signal intensity between spectra. A commonly used normalization method for mass spectrometric data is rescaling each spectrum by its total ion current, i.e. the area under the curve (AUC) (Fung and Enderwick, 2002; Sauve and Speed, 2004). Other common choices for the rescaling coefficient include the spectrum median or mean. Alternatively, choosing the average AUC over all spectra as the rescaling coefficient can do a global normalization. A global optimization assumes that the sample intensities are all related by a constant factor. This means that the data distribution should not differ substantially from one spectrum to another.

Peak detection deals with the selection of m/z values that display a reasonable intensity compared with those that display noise. Coombes et al. (2004) applied a simple peak finding (SPF) algorithm that provides the locations of potential peaks and their associated left-hand and right-hand bases. They estimated signal-to-noise ratio (S/N) using wavelets for improved peak detection. Also, they introduced a method for coalescing neighboring peaks.

Assuming that appropriate low-level analysis methods are used for mass spectral data preprocessing, biomarker selection can be addressed using various computational methods. One of the commonly used approaches is to apply statistical analyses that recognize differentially expressed m/z values between cases and controls with multiple subjects. For example, one can apply a two-sample t-test method to compare the protein intensities at each m/z value in cases and controls. Zhu et al. (2003) proposed a statistical algorithm that can select a subset of k biomarkers from the marker list that could best discriminate between the groups in a training dataset via the best k-subset discriminant method, with high sensitivity and specificity.

Machine learning methods have also been proposed for biomarker discovery. For example, Petricoin et al. (2002a) applied a combination of genetic algorithm (GA) and self-organizing clustering (GA-SOC) for variable selection. The GA-SOC, which is implemented in the ProteomeQuest software, starts with hundreds of random choices of small sets of exact m/z values selected from the SELDI-TOF mass spectra. Each candidate subset contains 5–20 of the potential m/z values that define the spectra. The m/z values within the highest rated sets are reshuffled to form new subset candidates. The candidates are rated iteratively until the set that fully discriminates the preliminary set emerges.

Koopmann et al. (2004) applied successfully support vector machines (SVMs) in a modified form to proteomic profiling. Li et al. (2002) introduced the unified maximum separability analysis (UMSA) algorithm, which incorporates data distribution information into structural risk minimization learning algorithm. UMSA is applied to identify a direction along which two classes of data are best separated. This direction is represented as a linear combination of the original variables. The weight assigned to each variable in this combination measures the contribution of the variable toward the separation of the two classes of data. They analyzed protein profiles of serum samples from the patient with or without breast cancer. They reported that UMSA enabled the identification of three discriminatory biomarkers that achieved 93% sensitivity and 91% specificity in detecting breast cancer patients from the non-cancer controls.

In our previous work (Ressom et al., 2005), we proposed a novel computational method known as PSO-SVM that combines SVMs and particle swarm optimization (PSO) for optimal selection of m/z values from high resolution SELDI-quadrupole-TOF (SELDI-QqTOF) spectra. First, we performed binning, normalization, baseline correction and peak identification. Then, we refined the identified peak list based on S/N of peaks and their frequency of occurrence in multiple spectra. Finally, we used the PSO-SVM algorithm to select optimal m/z values associated with the refined peak list.

In this paper, we performed peak alignment by combining neighboring peaks within a spectrum and across spectra. This peak alignment method defines windows of m/z values that have variable width. The PSO-SVM algorithm is applied to select the optimal m/z windows. We ran the algorithm multiple times and selected 7–9 m/z windows based on their frequency of occurrence. An SVM classifier that employs these m/z windows as its inputs yielded up to 91% sensitivity and 92% specificity in distinguishing hepatocellular carcinoma (HCC) patients from matched controls.

2 METHODS

2.1 Mass spectral data

The incidence of HCC in the United States is increasing. HCC has been associated with hepatitis C (HCV) and B (HBV) viral infections. Very high rates of HCC incidence are observed in Egypt, where an epidemic of viral infections presents a serious health problem. The management of the disease would benefit from identification of biomarkers related to this disease. Serum samples of HCC cases and controls were obtained from 2000 to 2002 in collaboration with the National Cancer Institute of Cairo University, Egypt. Controls were recruited among patients from the orthopedic fracture clinic at the Kasr El-Aini Hospital, Cairo, Egypt and were frequency-matched by gender, rural versus urban birthplace and age to cancer cases (Ezzat et al., 2005). Blood samples were collected by a trained phlebotomist each day around 10 am and processed within a few hours according to a standard protocol. Aliquots of sera for the mass spectrometric analysis were frozen at −80°C, immediately after collection till the analysis; all measurements were performed on samples of second-time thawed serum.

A total of 411 sera samples (199 from HCC patients and 212 from matched healthy individuals) were analyzed by using SELDI-QqTOF, a hybrid quadrupole time-of-flight (QSTAR, Applied Biosystems) mass spectrometer interfaced with a weak cation exchange (WCX) protein array (Ciphergen Biosystems). The protein array consists of eight spots at which samples are presented to the ionization source of the instrument. We used one of the spots for a reference serum and the rest for sera from cases and controls. Note that the same serum was used as a reference on each array throughout the study. To control the bias that maybe introduced by spot location, we analyzed 3 cases and 4 controls (or 3 controls and 4 cases) on each array. The spot locations were interchanged in consecutive
arrays. For example, if a spot location was used for a case in one array, it was used for a control or a reference in the next array. The spot location of the reference serum was determined on a rolling basis, i.e. it was moved from one to eight and back to one in each consecutive run.

The replicate spectra of the reference serum were used to assess technical variability. A total of 61 reference spectra were available for this study. Each spectrum had 6107 intensity values (found using a binning procedure discussed in the next section). We transformed each intensity value by computing the base-two logarithm and found the mean log intensity value and standard deviation. The coefficient of variation of the log-transformed intensity values in the 61 reference spectra ranged between 6.4% and 22.4%, with a mean value of 15.5%.

2.2 Low-level analysis

We applied low-level analysis methods to preprocess the raw high-resolution SELDI-QqTOF mass spectra. We began our analysis with outlier screening, where we removed spectra whose data distribution substantially deviated from others. To reduce the noise and dimensionality of the raw spectra, we used a binning procedure that divides the mz axis into intervals of desired length. The mean of the intensities within each interval was used as the protein expression variable in each bin. The low-frequency baseline of each spectrum was estimated by using multiple shifted windows of 200 bins. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum, yielding a baseline corrected spectrum. Each spectrum was normalized by dividing it by its total ion current. In addition, the spectra were scaled to have an overall maximum intensity of 100. For peak detection, a bin is identified as a peak if the sign of the intensity’s slope changes from positive to negative. Peaks with intensity below a predefined threshold-line were considered as noise and were discarded. To account for variation in the mz location (drifts) in different spectra, two peaks were coalesced if they differed in location by at most 2 bins or at most 0.08% relative mass. This method was based on the ideas of Eilers and Marx (1996) applied the method for baseline correction of 2D gel electrophoresis images. Furthermore, each spectrum was normalized by dividing it by its total ion current. Figure 1 depicts a SELDI-QqTOF mass spectrum of a healthy individual. On the horizontal axis are mz values and on the vertical axis are intensity measurements that indicate the relative ion abundance. The top figure is the raw spectrum and the bottom figure depicts the spectrum after binning. As shown in the figures, the binning algorithm has removed the high frequency noise, thus smoothing the spectrum. Also, it improved the alignment of the spectra (not shown). Figure 2 (top) depicts the regressed baseline of the spectrum. The baseline corrected and normalized spectrum is shown in Figure 2 (bottom).
The 176 case and 181 control spectra were split into training and testing datasets. The training dataset consisted of 200 samples (100 cases and 100 controls). The testing dataset had 157 samples (76 cases and 81 controls). We used the training dataset for peak detection and peak alignment. The training spectra were rescaled so that the maximum intensity across all spectra is 100. A bin is identified as a peak if a change in the sign of the intensity’s slope occurred. Note that peaks with intensity below a pre-defined threshold-line were considered as noise and were discarded. In this study, we used a threshold-line that decreases linearly from 1.5% of the overall maximum intensity to 0.1% as \( \frac{m}{z} \) increases from 1 to 11.5 kDa. Thus, intensity values below this threshold line were considered as noise. This threshold line decreases with an increasing \( \frac{m}{z} \) because of the observed decrease in the noise-level at higher \( \frac{m}{z} \) values.

To accommodate drifts introduced by the instrument, we aligned peaks by coalescing neighboring peaks within and across spectra into \( \frac{m}{z} \) windows. First, we selected peaks above a threshold-line that decreases linearly from 2.5 to 1%. Then, we combined these peaks if they differed in location by at most 2 bins or at most 0.08% relative mass. This step found 444 \( \frac{m}{z} \) windows in the training dataset. Following this, we considered peaks with intensities between the threshold-line that decreases from 2.5 to 1% and another threshold-line, which decreases from 1.5 to 0.1%. These peaks were added into previously identified \( \frac{m}{z} \) windows if they fell within 2 bins or at most 0.08% relative mass. Note that this step may increase the width of an \( \frac{m}{z} \) window if a peak is added from outside, otherwise the \( \frac{m}{z} \) window size remains unchanged except that the number of peaks in that window will increase. We retained \( \frac{m}{z} \) windows that consisted of peaks from at least five spectra and discarded the rest. This step resulted in 368 \( \frac{m}{z} \) windows that satisfied the criterion. Finally, we found the maximum intensity within the 368 \( \frac{m}{z} \) windows for each spectrum in the training dataset, yielding a 368 \( \times \) 200 training data matrix. The testing spectra were binned, baseline corrected and normalized in the same way as the training spectra. They were rescaled based on the parameters used to rescale the training spectra, so that the maximum intensity in the training dataset is 100. The 368 \( \frac{m}{z} \) windows defined by the training spectra were used to create a 368 \( \times \) 157 testing data matrix.

![Fig. 1. SELDI-QqTOF mass spectrum in the range between 1 and 11.5 kDa: raw (top figure) and binned (bottom figure).](image1)

![Fig. 2. SELDI-QqTOF normalized spectrum and regressed baseline (top figure) and baseline corrected (bottom figure). Note that the above spectra show intensity values in the range 0–50.](image2)

### 3.2 Biomarker selection using PSO-SVM

We used the PSO-SVM algorithm to select candidate biomarkers from the 368 peak-containing \( \frac{m}{z} \) windows. In this study, we arbitrarily targeted selection of 5 \( \frac{m}{z} \) windows. The algorithm began
with 50 particles, where each particle consisted of five randomly selected \( m/z \) values from the 368 windows (i.e. \( n = 5, N = 50 \) and \( L = 368 \)). A linear SVM classifier was built for each particle via the training dataset. The prediction power of each particle (5 \( m/z \) windows) was evaluated by measuring the performance of the SVM classifier in distinguishing the two classes through the \( k \)-fold cross validation and bootstrapping methods. We used \( k = 10 \) for this study. The most-fit particles contributed to the next generation of 50 candidate particles. This process continued until the performance of the SVM classifier converged or a pre-specified number of iterations was reached. The algorithm was repeated 350 times, 175 runs performed using the 10-fold cross validation method and 175 runs using the bootstrapping method. Figure 4 depicts the percentage of occurrence of \( m/z \) windows selected by the PSO-SVM. Note that the \( m/z \) windows are sorted in decreasing order of frequency and only the first 60 \( m/z \) windows are shown in the figure. As shown in the figure, the change in frequency of occurrence became small after the first 9 \( m/z \) windows. These \( m/z \) windows yielded 91% sensitivity and 88% specificity in distinguishing HCC patients from healthy individuals in the testing dataset.

We observed that the population size has no significant effect on feature selection. This is evident from the frequency plot shown in Figure 5, where we ran the PSO-SVM algorithm 350 times for \( N = 50 \) and \( N = 100 \). The frequency plot for this experiment yielded 8 potential markers with 91% sensitivity and 90% specificity. The first 7 markers are the same as those found in the previous experiment with \( n = 5 \) and \( N = 50 \). Figure 6 depicts these 7 \( m/z \) windows along with the training spectra, mean spectrum for cases and mean spectrum for controls.

For comparison, we applied a two-sample \( t \)-test method to identify differentially expressed \( m/z \) windows from the preprocessed spectra. The method selected 128 \( m/z \) windows out of 368 at the significance level \( p < 0.0001 \). An SVM trained with the 128 \( m/z \) windows yielded 90% sensitivity and 90% specificity in the independent dataset. The 8 \( m/z \) windows selected by the PSO-SVM in the above experiment were also selected by the \( t \)-test method as a part of 128 \( m/z \) windows. Each of these \( m/z \) windows is differentially expressed, with \( p < 0.0001 \), and has a fold change \( >2 \) between controls and cases. However, there were 122 \( m/z \) windows, with \( p < 0.0001 \), and 38 with fold change \( >2 \). This demonstrates the power of the PSO-SVM algorithm in identifying a small set of relevant candidate biomarkers despite the presence of large number of statistically significant potential candidates.

To examine any potential bias that may be introduced by parameter choice, we ran the PSO-SVM algorithm 600 times for various numbers of features (\( n = 5–10 \)) and particles (\( N = 50 \) and 100). Figure 7 depicts the frequency plot for this experiment. As shown in the figure, 7 out of the first 9 potential markers are the same as those found in the previous two experiments. These 9 \( m/z \) windows distinguished the HCC patients from healthy individuals in the testing dataset, with 91% sensitivity and 92% specificity.
To study the effect of preprocessing, we performed biomarker selection using spectra that were binned and normalized, but not baseline corrected. A total of 684 \( m/z \) windows were found from these spectra using our peak detection and alignment methods described earlier. The increase in the number of \( m/z \) windows is attributed to features that were not baseline corrected. The PSO-SVM algorithm was run 120 times for features ranging from \( n = 5 \) to \( 10 \), with \( N = 50 \) and 100 particles. The corresponding frequency plot provided 5 \( m/z \) windows, of which four were the same as those found in the above three experiments. These 5 \( m/z \) windows yielded 85% sensitivity and 90% specificity. This shows that baseline correction has an impact in improving the prediction accuracy.

Table 1 summarizes the improvement in classification performance, with baseline correction for various features (all bins, all \( m/z \) windows and the \( m/z \) windows were selected by the PSO-SVM algorithm when \( n \) was varied between 5 and 10).

<table>
<thead>
<tr>
<th>Features</th>
<th>Without baseline correction</th>
<th>With baseline correction</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sensitivity</td>
<td>Specificity</td>
</tr>
<tr>
<td>All ( m/z ) bins</td>
<td>91</td>
<td>85</td>
</tr>
<tr>
<td>All ( m/z ) windows</td>
<td>90</td>
<td>83</td>
</tr>
<tr>
<td>( m/z ) windows selected by PSO-SVM</td>
<td>85</td>
<td>90</td>
</tr>
</tbody>
</table>

4 CONCLUSIONS AND FUTURE WORK

In this paper, we presented low-level analysis methods for mass spectral data preprocessing. A computational method that combines particle swarm optimization with support vector machines is applied for biomarker selection. We showed that the proposed approaches can select mass points from the complex mass spectra. For the SELDI-QqTOF data presented in this paper, 7–9 \( m/z \) windows were selected that yielded up to 91% sensitivity and 92% specificity in distinguishing liver cancer patients from healthy individuals in an independent dataset. Compared with our previous study for the same data, we observed that the use of \( m/z \) windows provides equal or better performance than precise \( m/z \) values or \( m/z \) bins. The \( m/z \) windows selected by the PSO-SVM algorithm consist of clearly detectable peaks, which are more likely to represent identifiable proteins, protein fragments or peptides. This is important for our ultimate goal of identifying proteins/peptides that distinguish cancer patients from healthy individuals. Once the proteins are identified, focus will be on validating the proteins through other sample-sets and analytical platforms, such as the increasingly popular matrix-assisted laser-desorption/ionization time-of-flight (MALDI-TOF) mass spectrometry (Villanueva et al., 2004; Koomen et al., 2005).

We believe that the use of computational methods alone cannot provide a solution to the complex task of biomarker discovery from mass spectra involving thousands of proteins. In addition to advanced computational methods that are capable of extracting knowledge from complex and high dimensional data, this task requires a careful study design, sample collection and preparation, improved mass spectrometry, well-designed low-level analyses and inter-laboratory validation.

ACKNOWLEDGEMENTS

We thank J. Jelinek for the programming of the binning algorithm. We thank Drs F. Seillier-Moiseiwitsch and A. Wang for their helpful
suggestions and discussions. This work was supported in part by US Army Medical Research and Material Command, the Prostate Cancer Research Program grant DAMD17-02-1-0057 and the American Cancer Society grant CRTC-02-245-01-CCE awarded to R.G.

Conflict of Interest: none declared.

REFERENCES


