Particle Swarm Optimization for Analysis of Mass Spectral Serum Profiles

H. Ressom[1], R.S. Varghese[1], D. Saha[1], E. Orvisky[1], L. Goldman[1], E.F. Petricoin[2],
T.P. Conrads[3], T.D. Veenstra[3], M. Abdel-Hamid[4], C.A. Loffredo[1], and R. Goldman[1]

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

ABSTRACT

Serum profiling using mass spectrometry is an emerging technology with a great potential to provide biomarkers for complex diseases such as cancer. However, protein profiles obtained from current mass spectrometric technologies are characterized by their high dimensionality and complex spectra with substantial level of noise. These characteristics have generated challenges in discovery of proteins and protein-profiles that distinguish cancer patients from healthy individuals. This paper proposes a novel machine learning method that combines support vector machines with particle swarm optimization for biomarker discovery. Prior to applying the proposed biomarker selection algorithm, low-level analysis methods are used for smoothing, baseline correction, normalization, and peak detection. The proposed method is applied for biomarker discovery from serum mass spectral profiles of liver cancer patients and controls.

Categories and Subject Descriptors: I.5.2 [**Pattern Recognition**]: Design Methodology – *classifier design and evaluation, feature evaluation and selection.*

General Terms: Algorithms

Keywords: proteomics, support vector machines, swarm intelligence.

1. INTRODUCTION

Several laboratories have demonstrated the feasibility of using serum-based proteomic pattern analysis by mass spectrometry for biomarker discovery and diagnosis of ovarian [14], breast [13], and prostate cancer [1], [15]. One of the commonly used biomarker discovery approaches is to apply statistical analyses that recognize differentially expressed mass-per-charge (m/z) values between the two groups 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. [23] 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. [14] applied a combination of genetic algorithm (GA) and self-organizing clustering (GA-SOC) for variable selection. The GA-SOC, which is implemented in ProteomeQuest software, begins with a random generation of a population of many subsets of the surface-enhanced laser desorption ionization-quadrupole time of flight (SELDI-QqTOF) mass spectra with precise m/z candidate values. The user arbitrarily fixes the number of features (i.e., m/z values) that will create the best model. In their study, the number of features varies with the biologic state and ranges from 5 to 20.

One of the concerns in the construction and use of machine learning algorithms is the possibility of overfitting the training data [17]. Hence, it is necessary to have algorithms that are capable of not only dealing efficiently and effectively with high dimensionality of mass spectral data, but also producing models with good generalization capability. The latter demands models derived from a training dataset to apply equally well to a previously unseen (independent) dataset.

Support vector machines (SVMs), introduced by Vapnik [20], have proven themselves as one of the classification algorithms that have the potential to satisfy the above two demands. Parameters of SVMs are determined based on structural risk minimization. In classification problems that involve two linearly separable classes, SVMs search for one target known as the optimal hyperplane. The optimal hyperplane maximizes the margin of separation between the hyperplane and the closest data points on both sides of the hyperplane. Koopmann [12] applied successfully SVM in a modified form to proteomic profiling. While SVMs and their variants have been successfully used in classification tasks, the selection of the most salient m/z values from high dimensional mass spectrometric data remains a challenging task. To address this challenge, Li et al. [13] introduced 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 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.

Our proposed method takes advantage of the collective features of SVMs and particle swarm optimization (PSO). PSO is similar to

GECCO'05, June 25-29, 2005, Washington, DC, USA.
Copyright 2005 ACM 1-59593-010-8/05/0006…$5.00.

evolutionary computation methods such as genetic algorithms (GAs). Each uses a population of potential solutions to explore the search space. While GAs are based on survival-of-the-fittest approaches as in the theory of natural evolution, PSO is an adaptive algorithm based on the social metaphor of flocking birds (or schooling fish, or swarming insects). In PSO, a population of individuals adapt by stochastic search of successful regions of the search space, influenced by their own success and that of their neighbors. Individual particles move stochastically in the direction of their own previous best position and the best position discovered by the entire swarm. Alternatively a neighborhood approach can be used, where instead of moving in the direction of the best position discovered by the entire swarm, each particle moves towards the best position discovered amongst a localized group of particles, termed the "neighborhood." Since the change in particle trajectory is based on the position of the particle's own best position as well as the global (or neighborhood) best position, the essence of the PSO algorithm is that each particle will continuously focus and refocus the efforts of its search within these two regions. Each particle in the swarm represents a candidate solution to the optimization problem, and is evaluated at each update by a performance function. PSO is a simple algorithm that has been shown to perform well for optimization of a wide range of functions, often locating optima in difficult multi-modal search spaces faster than traditional optimization techniques. Detailed information on swarm intelligence and the PSO algorithm can be found in [3], [8], and [11]. We have successfully applied PSO in ocean color remote sensing application and showed that PSO requires less computation time than GA [18].

The paper is organized as follows: Section 2 highlights the methods for biomarker discovery. The section gives an overview of low-level analysis methods such as smoothing, baseline correction, normalization, and peak detection. In addition, the section introduces the proposed PSO-SVM algorithm and its application in biomarker discovery. Section 3 presents analyses made using our proposed biomarker discovery method and results obtained in analyzing liver cancer. Section 4 concludes the paper.

2. METHODS

2.1 Low Level Analysis

Analysis methods for biomarker discovery via mass spectrometric data will perform sub-optimally, if low-level analysis is not made properly. The reason for this includes the substantial amount of noise and systematic variations between spectra caused by varying amount of protein, sample degradation over time, and variation in the sensitivity of the instrument. Sorace and Zhan [19] have reported the existence of a significant non-biologic experimental bias between cancer and control subjects in their assessment of ovarian cancer serum proteomic profiling using SELDI-QqTOF. Unfiltered mass spectra contain electronic noise, chemical noise due to contaminants and the ionization matrix used, and protein signatures [15]. Previous quality-control experiments have suggested several measurement properties of current mass spectrometry technologies that must be accounted for in the analysis [9], [22]. 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.

Low-level analysis methods allow smoothing, baseline correction, and normalization of raw mass spectrum as well as peak detection. Smoothing can reduce the effect of some m/z values that appear as peaks but may not be or are very hard to verify by independent experiments. Baseline correction can minimize the effect of background noise. Normalization reduces systematic variation between spectra that may be caused by varying amounts of protein, sample degradation over time, or variation in the sensitivity of the mass spectrometry's ion detector. Peak detection deals with the selection of m/z values which display a reasonable intensity compared to those that appear as noise.

Many smoothing algorithms have been proposed in the field of signal processing to denoise raw signals including the well-known Savitzky-Golay filter that removes additive white noise. Pusztai et al. [16] have used this filter to process SELDI-QqTOF mass spectra. Wavelets are also useful in smoothing. For example, a multiresolution wavelet analysis can be used to estimate the approximation coefficients and the detail coefficients at different multiresolution levels. Smoothing can be achieved by thresholding the detail coefficients [7]. The spectra can be reconstructed by using the inverse wavelet transform without the thresholded detail coefficients. Several methods have been proposed for baseline correction. Baggely et al. [2] fitted a local median in a fixed window on the time scale. They also considered a local minimum instead of local median. Their final decision was to subtract a "semimonotonic" baseline. This procedure is described in [2]. A commonly used normalization for mass spectrometric data is rescaling each spectrum by its total ion current, i.e., the area under the curve (AUC). Alternatively, choosing the average AUC over all spectra as the rescaling coefficient can do a global normalization. Other common choices for the rescaling coefficient include the spectrum median or mean. A global optimization assumes that the sample intensities are all related by a constant factor. That means that the data distribution should not differ substantially from one spectrum to another.

Coombes et al. [5] applied a simple peak finding (SPF) algorithm to identify peaks. The algorithm provides the locations of potential peaks and their associated left-hand and right-hand bases. In [6], Coombes et al. introduced an improved peak detection method using wavelets. Through undecimated discrete wavelet transform, they separated the true signal from the noise. The noise component was used to estimate the average noise for each peak. The ratio of the baseline corrected intensity and the average noise at each peak (signal-to-noise ratio, S/N) was used to select reasonable peaks among those selected via the SPF. They selected peaks with S/N > 10. To accommodate some drift in the locations of spectral peaks from one experiment to another, two peaks are coalesced if they differed in location by at most 7 clock ticks or at most 0.3% relative mass for surface-enhanced laser desorption ionization-time of flight (SELDI-TOF) data. They also revisited peaks with 2 < S/N < 10 and added these to the list if they fell within the same distance limits (7 ticks or 0.3% of mass) of a previously identified peaks.

In this paper, we applied low-level analysis methods to raw high-resolution SELDI-QqTOF mass spectra. To reduce the noise and dimensionality of the raw spectra, we used a binning procedure that divides the m/z axis into intervals of desired length. The mean of the intensities within each interval was used as the protein expression variable in each bin [21]. The baseline of each spectrum was estimated by using multiple shifted windows of a given size. 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. For peak detection, we used the SPF algorithm and refined the peaks by choosing those with S/N >

2. Note that, in this study, we did not perform any alignment of the spectra or coalescing of peaks that are close to each other. We are investigating these approaches.

2.2 Biomarker Discovery

The purpose of this analysis is to identify biomarkers from the preprocessed mass spectral data. While peak detection deals with the selection of mass points with reasonable intensity and signal-to-noise ratio, the aim of biomarker discovery is to identify mass points that can be used to distinguish between cancer patients and health individuals.

Statistical analyses can be applied to recognize differentially expressed m/z values between the two groups from a high dimensional mass spectral datasets with multiple subjects. Alternatively, peak detection algorithms can be used to identify potential peaks and use the resulting peaks for further analysis. We believe that m/z values that may appear statistically insignificant or subtle peaks may still be useful in improving classification accuracy when they are used in combination with other m/z values. To extract m/z values that interact with each other and to have a more concise list of m/z values, advanced computational methods are needed.

In this paper, we propose the use of a PSO-SVM algorithm for biomarker discovery. The algorithm builds SVM classifiers for each particle (potential solution) generated by PSO. The prediction capability of the resulting SVM classifier on a validation dataset is used as a performance function for the PSO algorithm. Since SVMs provide good generalization capability in classification tasks and can be designed in a computationally efficient manner, they are an ideal candidate for use as a performance function.

2.2.1 Support Vector Machines

Support vector machines, introduced by Vapnik [20], are learning kernel-based systems that use a hypothesis space of linear functions in high dimensional feature spaces. Unlike artificial neural networks, which try to define complex functions in the input feature space, the kernel methods perform a nonlinear mapping of the complex data into high dimensional feature spaces and then use simple linear function to create linear decision boundaries. Thus, the problem of choosing network architecture is replaced here by the problem of choosing a suitable kernel for the data projection.

The advantages of support vector machines over neural networks is that they are significantly faster to train, better suited to work with high dimensional data, provide better generalization ability on the test set, can be developed with few training examples, and allow for scaling the importance of outliers. Parameters of SVMs are determined based on structural risk minimization.

In classification problems that involve two linearly separable classes (e.g. A and B in Figure 1), SVMs search for one target known as the optimal hyperplane. While various hyperplanes can separate the two groups correctly, the optimal hyperplane maximizes the margin of separation (ρ) between the hyperplane and the closest data points on both sides of the hyperplane. Thus, SVMs can be used to develop an optimal classifier that best separates cancer patients from healthy individuals based on m/z values.

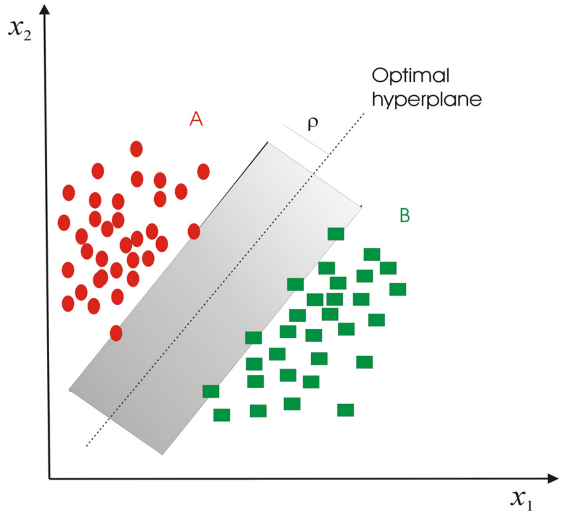


Figure 1: The optimal hyperplane that separates data points from two linearly separable classes.

2.2.2 Particle Swarm Optimization

In the PSO algorithm, each particle is represented as a *D*-dimensional vector $\vec{x}_i \in \Theta$, with a corresponding *D*-dimensional instantaneous trajectory vector $\Delta\vec{x}_i(t)$, describing its direction of motion in the search space at iteration *t*. The core of the PSO algorithm is the position update rule (1) which governs the movement of each of the *N* particles, $i=1,2,...N$, through the search space.

$$\vec{x}_i(t+1)=\vec{x}_i(t)+\Delta\vec{x}_i(t+1)$$
$$\Delta\vec{x}_i(t+1)=\chi\left(\Delta\vec{x}_i(t)+\Phi_1\left(\vec{x}_{i,best}(t)-\vec{x}_i(t)\right)+\Phi_2\left(\vec{x}_{G,best}(t)-\vec{x}_i(t)\right)\right) \quad (1)$$

where $\Phi_1=c_1\begin{bmatrix} r_{1,1} & 0 & 0 & 0 \\ 0 & r_{1,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{1,D} \end{bmatrix}$ and $\Phi_2=c_2\begin{bmatrix} r_{2,1} & 0 & 0 & 0 \\ 0 & r_{2,2} & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & r_{2,D} \end{bmatrix}$

At any instant, each particle is aware of its individual best position, $\vec{x}_{i,best}(t)$, as well as the best position of the entire swarm, $\vec{x}_{G,best}(t)$. The parameters c_1 and c_2 are constants that weight particle movement in the direction of the individual best positions and global best positions, respectively; and $r_{1,j}$ and $r_{2,j}$, $j=1,2,\ldots D$ are random scalars distributed uniformly between 0 and 1, providing the main stochastic component of the PSO algorithm. Figure 2 shows a vector diagram of the contributing terms of the PSO trajectory update. The new change in position, $\Delta\vec{x}_i(t+1)$, is the resultant of three contributing vectors: (i) the inertial component, $\Delta\vec{x}_i(t)$, (ii) movement in the direction of the global (or neighborhood) best, $\vec{x}_{G,best}$, and (iii) movement in the direction of individual best, $\vec{x}_{i,best}$.

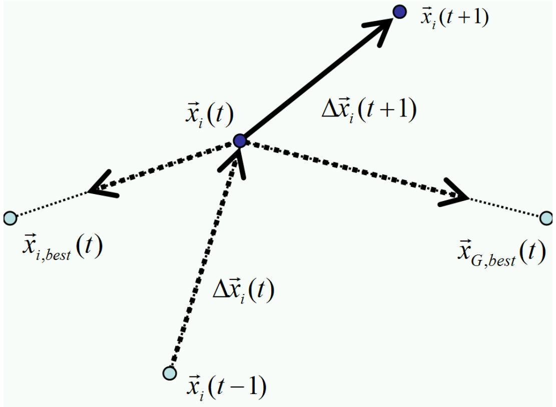


Figure 2: Vector diagram of particle trajectory update.

The constriction factor, χ, may also help to ensure convergence of the PSO algorithm, and is set according to the weights c_1 and c_2 as in (2) [3].

$$\chi = \frac{2}{\left|2-\varphi-\sqrt{\varphi^2-4\varphi}\right|}, \quad \varphi = c_1 + c_2, \quad \varphi > 4 \tag{2}$$

The key strength of the PSO algorithm is the interaction among particles. The second term in (1), $\Phi_2\left(\vec{x}_{G,best}(t)-\vec{x}_i(t)\right)$, is considered to be a "social influence" term. While this term tends to pull the particle towards the globally best solution, the first term, $\Phi_1\left(\vec{x}_{i,best}(t)-\vec{x}_i(t)\right)$, allows each particle to think for itself. The net combination is an algorithm with excellent trade-off between total swarm convergence, and each particle's capability for global exploration. Moreover, the relative contribution of the two terms is weighted stochastically.

The algorithm consists of repeated application of the velocity and position update rules presented above. Termination can occur by specification of a minimum error criterion, maximum number of iterations, or alternately when the position change of each particle is sufficiently small as to assume that each particle has converged. The PSO is implemented as a bounded optimization [10]. A pseudo-code description of the PSO algorithm is provided below:

(i) Generate initial population of particles, $\vec{x}_i$, $i = 1,2,\ldots N$, distributed randomly (uniform) within the specified bounds.

(ii) Evaluate each particle with objective function, $f(\vec{x}_i)$; if any particles have located new individual best positions, then replace previous individual best positions, $\vec{x}_{i,best}$, and keep track of the swarm global best position, $\vec{x}_{G,best}$.

(iii) Determine new trajectories, $\Delta\vec{x}_i(t+1)$, according to Eq. (1).

(iv) Update each particle position, $\vec{x}_i(t+1) = \vec{x}_i(t) + \Delta\vec{x}_i(t+1)$.

(v) Determine if any $\vec{x}_i(t+1)$ are outside of the specified bounds; hold positions of particles within the specified bounds.

(vi) If termination criterion is met (for example completed maximum number of iterations), then $\vec{x}_{G,best}$ is the best solution found; otherwise, go to step (ii).

Figure 3 demonstrates particle trajectories over a two-dimensional error surface. The contours represent equi-error curves on the error surface. Several of the depressions on the surface represent local minima (labeled "B"), whereas the point labeled "A" possesses the globally minimum error on the surface. A simple PSO with 10 particles was used to determine the position on this surface with minimum error, with their trajectories traced as lines. Note their mildly erratic behavior as each is pulled towards both the global and its individual best solution.

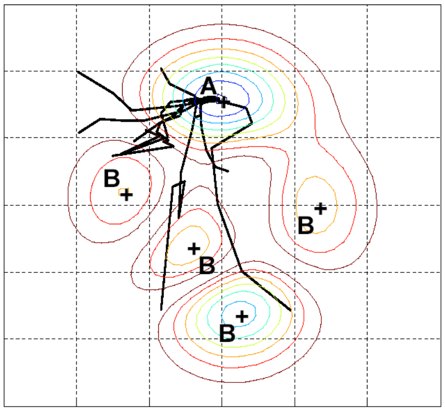


Figure 3: Trajectories of particles in solution-space.

2.2.3 PSO-SVM for Biomarker Discovery

Figure 4 depicts the proposed methodology for biomarker discovery. It starts by preprocessing the raw mass spectral data using the low-level analysis methods described in Section 2.1. The preprocessed data are split into training and testing (independent) datasets. The training dataset is used to select features and to build an SVM classifier. The validity of each classifier created with the selected features is evaluated using the sensitivity and specificity of the SVM classifier in distinguishing cancer patients from non-cancer controls. SVM classifiers are built for various combinations of features until the classification accuracy of the SVM classifier converges or maximum iteration number is reached. Estimates of classification accuracy are calculated by using the hold-out method where a validation dataset is used to evaluate the generalization error.

The PSO algorithm guides the selection of potential biomarkers that lead to best sensitivity and specificity in distinguishing cancer patients from healthy individuals. Figure 5 illustrates the PSO-SVM algorithm in more details. The ultimate biomarkers are evaluated using the testing (independent) dataset.

The PSO-SVM algorithm can be used to identify the optimal m/z values either from the entire variable set or from a reduced subset selected by other methods such as t-test or a peak detection algorithm. From these variables, the algorithm chooses *n* sets of randomly selected *k* m/z values (biomarkers) as initial particles. The algorithm evaluates the performance of each particle in distinguishing the two classes. This performance test is carried out by building an SVM classifier for each particle and using the cross-validation method. The algorithm uses the most-fit particles to contribute to the next generation of *n* candidate particles. Thus, on the average, each successive population of candidate particles fits better than its predecessor. This process continues until the performance of the SVM classifier converges. The algorithm repeats the above steps for various values of *k* to detect the optimal number of biomarkers along with the m/z values. The final biomarkers are evaluated via testing dataset (i.e., independent dataset that was not used for training) to determine the sensitivity and specificity of the SVM classifier.

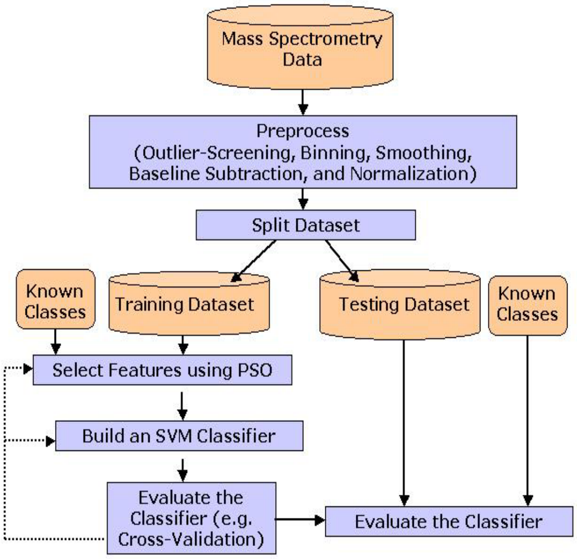

Mass Spectrometry Data
Preprocess (Outlier-Screening, Binning, Smoothing, Baseline Subtraction, and Normalization)
Split Dataset
Known Classes
Training Dataset
Testing Dataset
Known Classes
Select Features using PSO
Build an SVM Classifier
Evaluate the Classifier (e.g. Cross-Validation)
Evaluate the Classifier

Figure 4: Methodology for biomarker discovery.

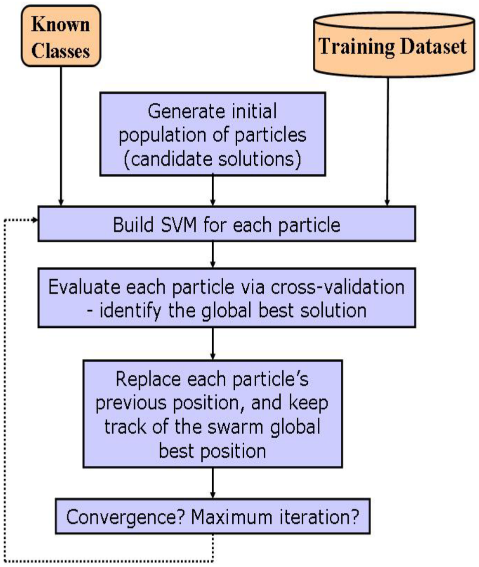

Known Classes
Training Dataset
Generate initial population of particles (candidate solutions)
Build SVM for each particle
Evaluate each particle via cross-validation - identify the global best solution
Replace each particle's previous position, and keep track of the swarm global best position
Convergence? Maximum iteration?

Figure 5: PSO-SVM algorithm.

3. RESULTS AND ANALYSIS

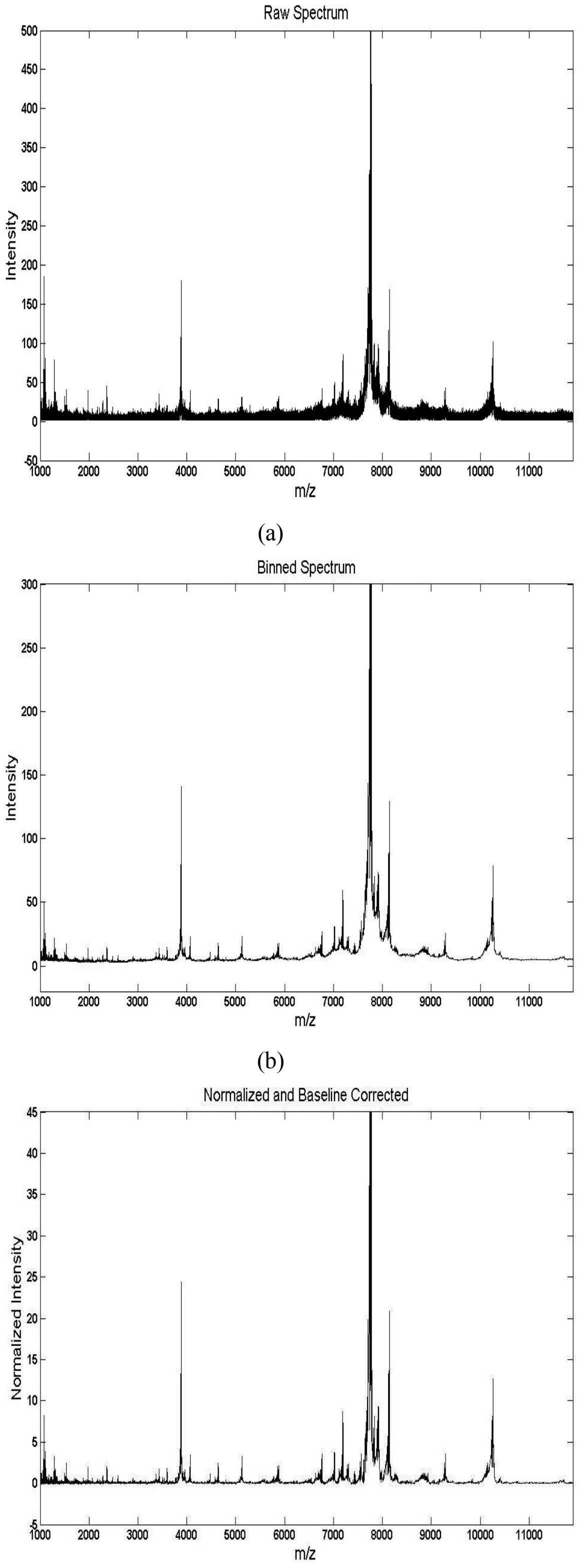
We examined the capability of the proposed PSO-SVM algorithm in identifying optimal m/z values (biomarkers) distinguishing liver cancer patients from healthy individuals. 411 SELDI-QqTOF mass spectra, 199 from hepatocellular carcinoma patients (cases) and 212 from matched healthy individuals (controls) were available from an ongoing study. About 13% of these spectra displayed substantial deviation from the data distribution and were excluded, leaving 357 (176 cases and 181 controls) spectra for further analysis. These outliers were singled out based on their deviation from the median ion current, median record count (number of mass points), and their alignment with pre-selected landmarks. Each spectrum consisted of ~340,000 m/z values with the corresponding ion intensities. The dimension of these high-resolution spectra was reduced to 6107 m/z values via a binning procedure that divides the m/z axis into intervals of desired length to the mass range 1 to 11.5 kDa. A bin size of 400 parts per million (ppm) was found adequate as it is 10 times the routine mass accuracy of the QqTOF with external calibration (40-50 ppm) [4]. The mean of the intensities within each interval was used as the protein expression variable in each bin [21]. The baseline of each spectrum was estimated by using multiple shifted windows of a given size. Spline approximation was used to regress the varying baseline. The regressed baseline was subtracted from the spectrum yielding a baseline corrected spectrum. Furthermore, each spectrum was normalized by dividing it by its total ion current.

Figure 6 depicts a typical SELDI-QqTOF mass serum spectrum of a healthy individual. On the horizontal axis are m/z values and on the vertical axis are intensity measurements that indicate the relative ion abundance. The raw spectrum is shown in Figure 6a. Figure 6b is the spectrum after binning. As shown in the figures, the binning algorithm has removed the high frequency noise, thus smoothing the spectrum. The binned spectrum is further normalized and baseline corrected (Figure 6c).

The preprocessed mass spectra were split into three datasets: training, validation, and testing datasets with 140 samples (70 cases and 70 controls), 60 samples (30 cases and 30 controls), and 157 samples (76 cases and 81 controls), respectively. We used the training and validation datasets (100 cases and 100 controls) for peak detection via the SPF method [5]. As we described in Section 2.1, an estimate of S/N is needed to select reliable peaks among those identified by the SPF method. To estimate the noise level in the preprocessed mass spectra, we applied undecimated discrete wavelet analysis that separated the true signal from noise. The noise for each peak was estimated by averaging over a window of 500 m/z values to the left of the peak and 500 m/z values to the right. We selected peaks with S/N > 2 and found 2940 peaks in the 200 spectra (100 cases and 100 controls). An SVM classifier trained with these 2940 peaks yielded 87% success in distinguishing liver cancer patients from healthy individuals in the testing dataset.

To reduce the number of peaks, we selected those peaks that were present in at least 25 of cases (out of 100) and combined them with those that were present in at least 25 of controls (out of 100). This approach resulted in 437 peaks. With these peaks, an SVM classifier achieved a prediction accuracy of 89% in distinguishing cases and controls in the testing dataset.

Further reduction of the number of peaks did not improve the prediction accuracy. For example, we found 59 peaks by combining peaks that were present in at least 50 cases with those that were present in 50 controls. These peaks resulted in only 83% prediction accuracy.



(a)

(b)

(c)

Figure 6: Typical SELDI-QqTOF mass spectrum in the range between 1-11.5 kDa: raw (a), binned (b), and normalized and baseline corrected (c).

Figure 7 and 8 depict the averaged spectrum for the 100 controls and 100 cases, respectively. These figures indicate 384 peaks that were present in at least 25 controls and 394 peaks found in at least 25 cases, respectively. Figure 9 shows the absolute difference between the averaged control spectrum and averaged case spectrum along with the 437 peaks found by combining those that were present in at least 25 cases with those that were present in at least 25 controls. The reduction in the combined number of peaks (i.e., 384+394 > 437) indicates overlaps between case and control peaks.

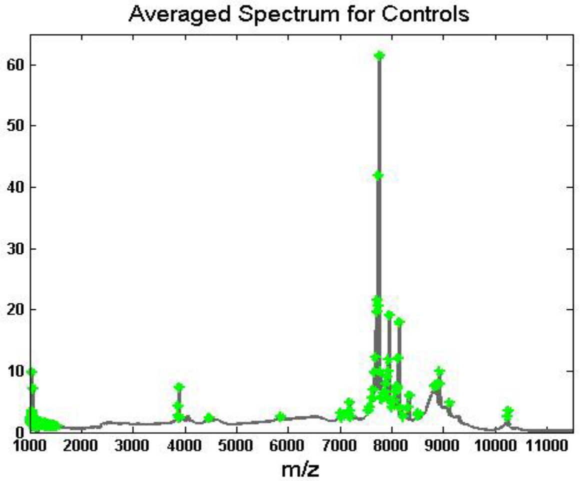


Figure 7: Averaged spectrum and peaks present in at least 25% of the control spectra.

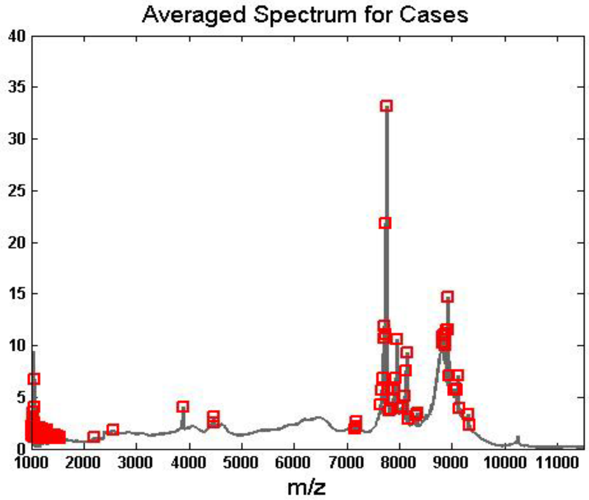


Figure 8: Averaged spectrum and peaks present in at least 25% of the case spectra.

We used the PSO-SVM algorithm to select biomarkers from the peaks that we identified (above). In this study, we arbitrarily targeted at five biomarkers. The parameters of the PSO algorithm were selected as $c_1 = c_2 = 2.05$ (Eq. 1), thus $\chi = 0.73$ (Eq. 2). The algorithm began with 100 particles where each particle consisted of k randomly selected m/z values from the 437 peaks. A linear SVM classifier was built for each particle via the training dataset. The prediction power of each particle (k biomarkers) was evaluated. This was done by measuring the performance of the SVM classifier in distinguishing the two classes in the validation dataset. The most-

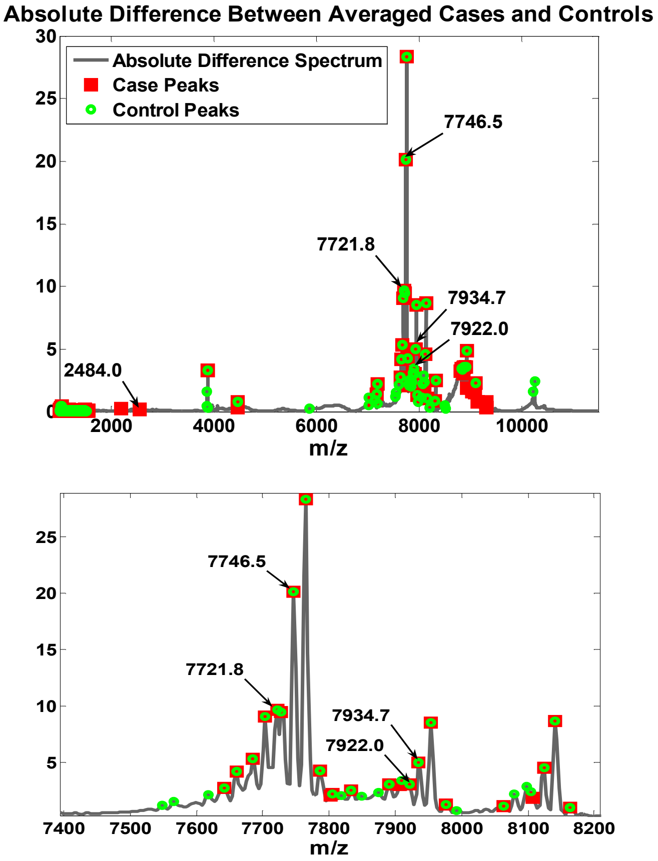


Figure 9: Absolute difference between averaged cases and controls; peaks present in at least 25% of cases; peaks present in at least 25% of controls; biomarkers found by PSO-SVM. Mass range: 1-11.5KDa (top figure); 7.4-8.2KDa (bottom).

fit particles contributed to the next generation of 100 candidate particles. This process continued until the performance of the SVM classifier converged. The algorithm was run for *k* = 3, 5, 10, 15, 20, and 25. We selected five biomarkers (7721.8, 7746.5, 7922.0, 7934.6, and 8331.4) that were identified by the algorithm more frequently than other peaks. These biomarkers yielded 91% prediction accuracy (88% sensitivity and 94% specificity) in distinguishing liver cancer patients from the non-cancer controls in the testing dataset. Note that the testing dataset consisted of 76 cases and 81 controls that were used neither for training nor for biomarker selection. We also applied the PSO-SVM algorithm to select biomarkers from the 2940 peaks that were present at least once in anyone of the case or control spectra. The algorithm was run for the same values of *k* as above. We selected five biomarkers (2484.0, 7721.8, 7746.5, 7922.0, and 7934.6) that were identified by the algorithm more frequently than other peaks. These peaks resulted in 92% prediction accuracy (91% sensitivity and 93% specificity) in detecting liver cancer patients from the non-cancer controls in the testing dataset.

To examine the performance of the PSO-SVM algorithm further, we ran the algorithm assuming all 6107 m/z values as potential biomarkers. The algorithm was run for the same values of *k* as in the above analyses. Five biomarkers (2484.0, 7721.8, 7746.5, 7922.0, and 8328.0) were found more frequently than other m/z values. These biomarkers yielded 92% prediction accuracy (90% sensitivity and 95% specificity) on the testing dataset. This exercise demonstrates the power of the algorithm in identifying relevant biomarkers despite the presence of large number of unlikely peaks.

Finally, we selected five biomarkers (2484.0, 7721.8, 7746.5, 7922.0, and 7934.6) that appeared frequently among the seven biomarkers found in the above three analyses. These biomarkers resulted in 92% prediction accuracy (91% sensitivity and 93% specificity). Figure 9 presents those five biomarkers. As shown in the figure, PSO-SVM selected not only obvious peaks but also subtle peaks, which appear insignificant. This demonstrates that the algorithm is capable of identifying peaks whose interaction with other peaks leads to more accurate classification. Table 1 summarizes the results obtained in this study.

Table 1: Five biomarkers selected from various variable sets

Variable Set	Biomarkers Selected	Sensitivity	Specificity
6107	2484.0, 7721.8, 7746.5, 7922.0, 8328.0	90	95
2940	2484.0, 7721.8, 7746.5, 7922.0, 7934.6	91	93
437	7721.8, 7746.5, 7922.0,7934.6, 8331.4	88	94
7	2484.0, 7721.8, 7746.5, 7922.0, 7934.6	91	93

4. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a new computational method that combines particle swarm optimization with support vector machines (PSO-SVM) for biomarker discovery. We showed that PSO-SVM can select relevant m/z values from the complex mass spectra. The mass points selected by PSO-SVM provide improved prediction accuracy in blinded dataset compared to a predictor that uses all potential mass points or peaks. PSO-SVM has the capability to identify significant biomarkers despite the presence of large number of peaks with signal only slightly higher than noise. Selection of peaks with a reasonable signal-to-noise ratio can slightly improve the performance of the algorithm for biomarker selection.

We believe that further improvement in biomarker discovery can be achieved by optimizing the low-level analysis and the PSO-SVM algorithm. Thus, our future work will focus on two major tasks. First, we will continue to investigate low-level analysis methods for smoothing, baseline correction, normalization, peak detection, and alignment. Second, we will optimize the parameters of the PSO-SVM algorithm to improve its performance. The parameters include number of particles, number of iterations, values for the stochastic component of the PSO algorithm, appropriate kernels (e.g. linear, polynomial, radial basis, etc.) for SVM, number of biomarkers, and fitness measure to evaluate the performance of potential biomarkers.

We also 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 careful study design, sample collection and preparation, improved mass spectrometry, well-designed low-level analyses, and inter-laboratory validation.

5. ACKNOWLEDGMENT

This work was supported in part by U.S. Army Medical Research and Material Command, Prostate Cancer Research Program grant DAMD17-02-1-0057 and American Cancer Society grant CRTG-02-245-01-CCE awarded to RG.

6. REFERENCES

[1] Adam, B.L., Qu, Y., Davis, J.W., Ward MD, Clements, M.A., Cazares, L,H,, Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z., and Wright, G.L. Jr. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res., 62, 13,* (2002), 3609-3614.

[2] Baggerly, K.A., Morris, J.S., Wang, J., Gold, D., Xiao, L.C., and Coombes, K.R. A comprehensive approach to the analysis of MALDI-TOF proteomics spectra from serum samples. *Proteomics, 3, 9,* (2003): 1667-1672.

[3] Clerc, M. and Kennedy, J. The particle swarm: explosion stability and convergence in a multi-dimensional complex space. *IEEE Trans. Evolutionary Computing,* 6, 1, (2002) 58-73.

[4] Conrads, T.P., Fusaro, V.A., Ross, S., Johann, D., Rajapakse, V., Hitt, B.A., Steinberg, S.M., Kohn, E.C., Fishman, D.A., Whitely, G., Barrett, J.C., Liotta, L.A., Petricoin, E.F. 3rd, and Veenstra, T.D. High-resolution serum proteomic features for ovarian cancer detection. *Endocr Relat Cancer*. 11, 2, (2004), 163-178.

[5] Coombes, K.R., Fritsche, H.A. Jr, Clarke, C., Chen, J.N., Baggerly, K., A., Morris, J.S., Xiao, L.C., Hung, M.C., and Kuerer, H.M. Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization. *Clin Chem. 49,10* (2003), 1615-1623.

[6] Coombes, K.R., Tsavachidis, S., Morris, J.S., Baggerly, K.A., Hung, M.C., Kuerer, H.M. *Improved peak detection and quantification of mass spectrometry data acquired from surface-enhanced laser desorption and ionization by denoising spectra with the undecimated discrete wavelet transform*. Technical Report UTMDABTR-001-04, The University of Texas M.D. Anderson Cancer Center, 2004.

[7] Donoho, D.L. and Johnstone, I.M. Adapting to Unknown Smoothness via Wavelet Shrinkage, *Journal of the American Statistical Association, 90,* (1995), 1200-1224.

[8] Engelbrecht A.P. *Computational Intelligence: An Introduction*. Wiley, England, 2003.

[9] Fung, E.T. and Enderwick, C. ProteinChip clinical proteomics: computational challenges and solutions. *Biotechniques*, *32*(suppl) (2002), 34-41.

[10] Hu X, Eberhart R, and Shi Y. Engineering optimization with particle swarm, *IEEE Swarm Intelligence Symposium*, Indianapolis, IN, 2003.

[11] Kennedy, J. and Eberhart, R.C. Particle swarm optimization. In *Proceedings of the 1995 IEEE International Conference on Neural Networks* (Perth, Australia), vol. IV, pp. 1942–1948.

[12] Koopmann, J., Zhang, Z., White, N., Rosenzweig, J., Fedarko, N., Jagannath, S., Canto, M.I., Yeo, C.J., Chan, D.W., Goggins, M. Serum diagnosis of pancreatic adenocarcinoma using surface-enhanced laser desorption and ionization mass spectrometry. *Clin Cancer Res.*, *10*, *3*, (2004), 860-868.

[13] Li, J., Zhang, Z., Rosenzweig, J., Wang, Y.Y., Chan, D.W. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem. 48, 8*, (2002), 1296-1304.

[14] Petricoin, E.F., Ardekani, A.M., Hitt, B.A., Levine, P.J., Fusaro, V.A., Steinberg, S.M., Mills, G.B., Simone, C., Fishman, D.A., Kohn, E.C., Liotta, L.A. Use of proteomic patterns in serum to identify ovarian cancer, *Lancet, 359*, (2002), 572-577.

[15] Petricoin, E.F., Ornstein, D.K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velassco, A., Trucco, C., Wiegand, L., Wood, K., Simone, C.B., Levine, P.J., Linehan, W.M., Emmert-Buck, M.R., Steinberg, S.M., Kohn, E.C., Liotta, L.A. Serum proteomic patterns for detection of prostate cancer. *J Natl Cancer Inst., 94, 20,* (2002), 1576-1578.

[16] Pusztai, L., Gregory, B.W., Baggerly, K.A., Peng, B., Koomen, J., Kuerer, H.M., Esteva, F.J., Symmans, W.F., Wagner, P., Hortobagyi, G.N., Laronga, C., Semmes, O.J., Wright, G.L. Jr, Drake, R.R., and Vlahou, A. Pharmacoproteomic analysis of prechemotherapy and postchemotherapy plasma samples from patients receiving neoadjuvant or adjuvant chemotherapy for breast carcinoma. *Cancer, 100, 9,* (2004), 1814-1822.

[17] Qu, Y., Adam, B.L., Yasui, Y., Ward, M.D., Cazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J., Wright, G.L. Jr. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem., 48, 10,* (2002*),* 1835-1843.

[18] Slade, W.H., Ressom, H.W., Musavi, M.T., Miller, R.L. Inversion of ocean color observations using particle swarm optimization. *IEEE Transactions on Geoscience and Remote Sensing, 42, Issue 9,* (2004), 1915-1923.

[19] Sorace, J.M. and Zhan, M. A data review and re-assessment of ovarian cancer serum proteomic profiling. *BMC Bioinformatics 4, 1*, (2003), 24.

[20] Vapnik, V. *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.

[21] Villanueva, J., Philip, J., Entenberg, D., Chaparro, C.A., Tanwar, M.K., Holland, E.C., Tempst, P. Serum peptide profiling by magnetic particle-assisted, automated sample processing and MALDI-TOF mass spectrometry. *Anal Chem. 76, 6*, (2004), 1560-1570.

[22] Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.-L., Wright, G.L. Jr., Qu, Y., Potter, J.D., Winget, M., Thornquist, M. and Feng, Z. A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection. *Biostatistics, 4*, (2003), 449-463.

[23] Zhu, W., Wang, X., Ma, Y., Rao, M., Glimm, J., and Kovach, J.S. Detection of cancer-specific markers amid massive mass spectral data. *Proc Natl Acad Sci USA, 100, 25*, (2003), 14666-14671.