Abstract—Pathologic states within the prostate may be reflected by changes in serum proteomic patterns. Mass spectrometry is becoming an important tool that generates the proteomic patterns. Mass spectrometry yields complex functional data for which the features of scientific interest are the peaks. Due to this complexity of data, a higher order analysis such as wavelet transform is needed to uncover the differences in proteomic patterns. We have applied wavelet based feature extraction method to available data and used a filter approach to feature subset selection in order to identify the appropriate biomarkers from reconstructed mass spectra. Using different classification algorithms, our approach yielded an accuracy of 95%, specificity of 95%, and sensitivity of 96%.

Keywords: Proteomics, Cancer diagnosis, Wavelet transform, Classification, Biomarker.

I. INTRODUCTION

The development of tools for the early cancer diagnosis is a major problem, and clinicians have investigated a variety of diagnosis techniques. Recently, they have discovered that pathological changes within an organ might be reflected in proteomic patterns in serum [1]. The word ‘proteome’ coined in 1994, designates the complete set of proteins that ultimately results from genome transcription in a given cell, tissue, or organism [2]. Hence, Proteomics is the science of making qualitative and quantitative comparisons and contrasts of proteomes under different conditions (normal vs. cancer, treated vs. untreated) to understand biological processes (disease). The field of proteomics has since evolved to include almost any type of technology that focuses upon the wide-scale analysis of proteins [3][4].

The mass spectrometry is a tool, which provides information about proteins and their fragments. The definition of a mass spectrometer may seem simple: it is an instrument that can ionize a sample and measure the mass-to-charge ratio of the resulting ions. Therefore, mass spectrometer can give qualitative and quantitative information on the elemental, isotopic, and molecular composition of organic samples [5]. The mass spectrum analysis is a fast inexpensive procedure based on a sample of patient’s blood, and it may potentially allow cancer screening without any complication in time of diagnosis.

Prostate cancer is the most common cancer and the second leading cause of cancer-related death among men of very countries [6]. Application of mass spectrometry for diagnosis of prostate cancer could have an important effect on public health, but to achieve this goal new biomarkers are essential. Since most prostate cancers show few symptoms in their early stages, the disease is often diagnosed in an advanced stage. Prostate-specific antigen (PSA) have proven a useful biomarker for the detection of cancer, unfortunately the specificity of elevated PSA levels in this cohort is 25%-35%; therefore, 70%-75% of men undergoing biopsy because of an abnormal PSA level do not have prostate cancer [7]. These limitations of the PSA test call for efforts toward discovery of biomarkers more effective for diagnosis and prognosis of prostate cancer. Hence, the proteomic pattern analysis may be used in the future to aid clinicians so that fewer men are subjected to unnecessary biopsies.

From a modeling viewpoint, the mass spectra can be considered complex functional data in which the key features of scientific interest are the peaks [8]. The peaks represent proteins or protein fragments (peptides) in proteomic pattern. Raw mass spectrometry data tends to be incomplete, noisy, high correlation within the spectrum profile, high dimensional, etc. and hence not directly suitable for feature extraction. Additionally, mass spectral data display variations in the protein profile even at identical instrumental settings and sample conditions. In order to minimize the effect of irrelevant sources of variations such as humidity, time, etc. and to be able to more fully extract the information of mass spectra, a more sophisticated preprocessing method that de-noises as well as compresses needs to be utilized.

The wavelet transform (WT) is an effective tool for dimension reduction and noise removal in the analysis of proteomic data. Wavelets are very popular in signal processing because they are able to analyze both local and global behavior of functions. The WT is a projection of the spectrum onto an orthogonal basis, called a wavelet basis.
This is to say that the spectrum can be represented by a set of localized orthogonal basis functions called wavelets. Thus, wavelet analysis could provide de-noised and compressed representation of mass spectrometry data that make the feature extraction process more efficient and accurate due to many favorable properties, such as hierarchical and multiresolution decomposition structure, de-correlated coefficients, and a wide variety of orthogonal basis function possibilities.

We have applied wavelet-based feature extraction method to the mass spectra of prostate cancer patients and those of healthy people. We have used a filter approach for feature subset selection. We have employed the reconstructed mass spectra to identify the appropriate biomarkers and to evaluate the classification performance. Our results have confirmed that the mass spectrometry proteomic profiles allow the diagnosis of prostate cancer. Therefore, the wavelet-based reconstructed mass spectra can be a viable method in diagnosis of prostate cancer. For our developed technique, the accuracy was 95% on the dataset, its specificity was 94.9%, and its sensitivity was 95.6%. The paper is organized as follows: section II describes the materials and dataset that we use in this research and the preprocessing that we apply on mass spectrometry data; section III explains which steps of feature extraction and subset selection method is used; in section IV we investigate the effect of our approach on protein selection and classification performance; finally we conclude in section V.

II. DATA AND PREPROCESSING

We have used the surface enhanced laser desorption-ionization time-of-flight (SELDI-TOF) mass spectrometry (MS) serum proteomic patterns as the input data to our processing algorithms. Serum SELDI-TOF mass spectra data were then used for screening patients and a healthy population.

A. Dataset

The serum SELDI-TOF MS dataset were used in this research to identify serum proteomic patterns that differentiate the serum of prostate cancer from non-cancer controls cases. The dataset were downloaded from the freely available datasets of the American National Cancer Institute (NIC). As explained on the website, Dataset was collected using the WCX2 protein chip, and includes 322 samples which has been divided into 190 with benign prostate hyperplasia, 63 controls, and 69 with malignant prostate cancer. The distribution of samples for prostate cancer dataset is represented in Table I.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Num. of Cancer</th>
<th>Num. of Control</th>
<th>Num. of Benign</th>
</tr>
</thead>
<tbody>
<tr>
<td>Prostate 7-3-02</td>
<td>69</td>
<td>63</td>
<td>190</td>
</tr>
</tbody>
</table>

Briefly when a biological sample is submitted to mass spectrometer it is applied on a surface mixed with an energy absorbing matrix that makes it crystallize as it dries. The surface is then placed into a vacuum chamber and hit by a laser. The matrix absorbs energy and transfers it to the molecules of biological sample; as a result proteins desorb and ionize. Next a brief electric field is applied which accelerates the ionized proteins into a flight tube where they drift until they strike a detector that records the time of flight. Given the length of the tube and the applied voltage a quadratic transformation is used to drive the mass-to-charge ratio of the protein from the time of flight [9].

A mass spectrum is represented by a curve where the x-axis indicates the ratio of the weight of a specific molecule to its electrical charge (M/Z, in Daltons per unit charge) and the y-axis is the signal intensity for the same molecule as a measure of the abundance of that molecule in the sample. Each mass spectrum curve represents the expression profile of 15154 peptides defined by their M/Z ratios with corresponding intensities. Fig. 1 shows three examples of serum mass spectrum data, which are from a prostate cancer, a control case, and a benign hyperplasia respectively. All samples in dataset are divided into cancer and non-cancer (including control and benign cases) classes in this study. Criticisms have been raised over the validity of the information used in this research [10].

Previous works on this data include [4] who, on a subset of the data ran genetic algorithms classifier to differentiate between 2 groups (control, cancer), training on 56 observations (25, 31) and testing on 266 observations (228, 38).
groups and obtained 95% sensitivity and 78% specificity. On a different SELDI-TOF dataset used Adaboost and boosted decision trees and stumps also to distinguish between two patient disease status; control and prostate cancer [11]. Using training set of 74 observations (30, 44) and testing on 88 observations (28, 60) achieved sensitivity and specificity of 93.8%. These combined results from other researchers suggest that SELDI-TOF MS profiles can be used to differentiate the control (benign prostate hyperplasia and non-cancerous) and prostate cancerous status of patients.

B. Preprocessing

The following conceptual model could consider for mass spectral data. Suppose we observed N spectra, each taken on the same equally-spaced grid of length T of TOFs tj, j=1, ..., T. A model for y_i(t_j), the observed spectral intensity for spectrum i at TOFs t_j is

\[ y_i(t_j) = B_i(t_j) + N_iS_i(t_j) + \epsilon_{ij} \]  

(1)

The true signal, S_i(t), consist of a sum of possibly peaks, each corresponding to a particular biological molecule. The normalization factor, N_i, is a constant multiplicative factor to adjust for spectrum-specific variability. The baseline function, B_i, represents a systematic artifact commonly seen in mass spectrometry data. This artifact is believed to be attributable to a cloud of matrix molecules hitting the detector in the early part of the experiment, or to detector overload [12]. The electrical noise, \( \epsilon_{ij} \), we assume that is zero-mean Gaussians with the variance a function of time.

According to above model, certain preprocessing steps must be performed before analyzing the spectra, including removal of baseline, noise elimination and normalization to calibrate the spectra from different samples. We performed baseline correction on all spectra by using a nonlinear filter known as the “top-hat” procedure [13]. The normalization is done via total ion current (TIC) method, which is equivalent to the normalization with the L1 norm of spectrum. Noise removal was done instead on wavelet coefficients. Fig. 2 shows a typical mass spectrum with the baseline and the same processed spectrum without the baseline.

III. METHODS

A. Feature Extraction

As described in the previous sections, a mass spectrum is very spiky functional data that the key features are the peaks. Hence, the analysis method must be considering both local and global behavior of the mass spectra and extract the signal-pattern features. This fact motivates using a feature extraction method such as wavelets to account for this potentially useful information.

Wavelets are very popular in signal processing because they are able to analyze both local and global behavior of functions. The mass spectra are a non-stationary signal and thus wavelet is a suitable tool for preprocessing of this type of data. A mass spectrum can be considered as a sampling of a function \( f(x) \) with a given resolution N the number of points-features used to describe the spectrum.

Wavelets are families of orthonormal basis functions that can be used to parsimoniously represent other functions. For example, in \( L^2(\mathbb{R}) \), an orthogonal wavelet basis is obtained by dilating and translating a mother wavelet \( \Psi \) as \( \Psi_{jk}(x) = 2^{j/2}\Psi(2^j x - k) \) with \( j,k \) being integers. A function \( f \) can then be represented by the wavelet series, as follows

\[ f(x) = \langle f, \Phi \rangle \Phi(x) + \sum_{j=0}^{J-1} \sum_{k=0}^{2^j-1} \langle f, \Psi_{jk} \rangle \Psi_{jk}(x) \]  

(2)

The set \( \{ \langle f, \Phi \rangle, \langle f, \Psi_{jk} \rangle \} \) for \( j=0, \ldots, 2^J, \) \( k=0, \ldots, 2^j-1 \) is the set of wavelet coefficients. As we ascend in the level of detail, increasing \( j \), wavelet coefficients become smaller and smaller except from parts of the signal where spiky behavior is observed. By thresholding the wavelet coefficients we can reconstruct a de-noised version of the signal retaining the regions in which peaks are present [14]. The threshold determines the number of finally detected peaks, higher values of that threshold lead to lower numbers of detected peaks. We should try for a careful balance; as values of the wavelet threshold become higher and higher we do not remove only noise but we also start removing a part of the signal that potentially contains valuable discriminatory information. However, the select of suitable threshold is critical for effective denoising of mass spectrum signal. In Fig. 3, we
give an example of the noisy mass spectra and the denoised smoothed signal on a highly zoomed interval of the spectrum given in Fig. 1. It is exactly the threshold of the wavelet coefficients that controls the dimensionality of the finally produced feature space after the peak detection. Large values of the threshold result in fewer detected peaks and thus lower dimensionality.

B. Feature Selection

The feature selection methods are generally classified into two categories: filter and wrapper methods. In filter methods, the feature selector is independent of the specific learning algorithm used in classification and is used as a filter to discard irrelevant and/or redundant features. On the other hand, in wrapper methods, the feature selector behaves as a wrapper around the specific learning algorithm depending on which relevant features are determined [15].

In this study, we use the filter approach to feature selection in the mapped space of mass spectrum data. For this, we implement a statistical testing using a distance measure, called "Fisher’s criteria", defined as follows

\[ FC = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \]  

(3)

Where \( \mu_1 \) and \( \mu_2 \) are the arithmetic means for the wavelet coefficients at each point of the cancer and non-cancer groups, respectively. \( \sigma_1 \) and \( \sigma_2 \) are the standard deviation of the corresponding coefficients at each point for both cancer and non-cancer groups.

IV. RESULTS

We applied our approach for prostate cancer detection using serum SELDI-TOF MS data, as described in section II. As wavelet basis, we chose Daubechies [16], which has been reported previously to have a good performance on this field [17]. As a result, we considered an appropriate distance window to adequately differentiate various peaks corresponding to different molecules between selected points, which prevents the choice of points with correlated values [18][19].

We used three standard measures of the effectiveness of diagnosis technique: sensitivity, specificity, and accuracy. The sensitivity is the probability of the correct diagnosis for a patient with cancer, the specificity is the chances of the correct diagnosis for healthy person, and the accuracy is the probability of the correct diagnosis for the overall population of healthy and sick people. For evaluation of performance of our approach, we have applied linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), and neural networks to the reconstructed mass spectra of prostate cancer patients and healthy people.

We obtained wavelet transform of each mass spectrum in level, \( J = 8 \), with Daubechies four mother wavelet. Then we apply a threshold value to wavelet coefficients for denoising of mass spectra data in which the suitable threshold value was chose that lead to highest accuracy in dataset. Using Fisher’s criteria, the twenty coefficients were selected from each approximation and details in the wavelet space. The other coefficients were essentially zero. Then, the processed mass spectrum of each sample was reconstructed from residual coefficients. Fig. 4 shows two examples of reconstructed mass spectrum, which are from a healthy and a prostate cancer case respectively.

After reconstruction, the processed mass spectrometry

![Fig. 3. SELDI-TOF MS de-noised sample plots. (a) a noisy and (b) a de-noised spectrum.](image)

![Fig. 4. SELDI-TOF MS reconstructed sample plots. (a) a control and (b) a prostate case.](image)
raw data used to determine the differences between the prostate cancer and healthy people. Because this is the detection task, the processed serum samples in dataset were divided into cancer and non-cancer groups in which the control and benign samples were grouped as a control set. Then, we split dataset into training and testing using random permuting method. Half of samples in each group were used for the training and the remaining for the testing. Again, we applied the feature selection method to processed mass spectra for extraction of significant M/Z points. We implemented an experimental procedure that allowed us to control the number of M/Z points and minimal distance between selected points. We determined the number of points and minimal distance between points that lead to the highest accuracy in dataset. The optimal number of selected points was determined to be fifteen and the minimum distance thirty-two.

Table II shows the results for correctly classified mass spectra from the testing set for dataset described in Table I. For each classification algorithms, the accuracy, specificity, and sensitivity calculated and the results of detection on dataset with three different classifiers are listed in Table III. The results show that all three classification techniques reach the same accuracy, specificity, and sensitivity for studied dataset with selected M/Z points. However, we have shown that the selected biomarkers could discriminate fairly the cancer case from non-cancer patients using proteomic patterns.

We identify a set of 15 M/Z ratios by our approach for the wavelet-based reconstructed mass spectra that the peptides are (497.5, 504.6, 515.2, 478.5, 433.9, 1104.5, 127.1, 883, 1282.3, 122.9, 929.6, 863.7, 115, 420.8 and 1115). Our results agree with the results obtained by other previously published works [20]. We have shown that the wavelet-based reconstructed mass spectra is a viable approach which can be used to diagnosis of prostate cancer from proteomic patterns.

V. CONCLUSIONS

We studied the problem of diagnosing prostate cancer by analysis of mass spectrum data. We applied a wavelet-based preprocessing method to available dataset. We used a filter approach for selection of coefficients in the wavelet space, and reconstructed the processed mass spectra from the chose coefficients. We applied LDA, QDA, and neural networks (N.N.) to determine the effectiveness of these points in prostate cancer detection. For available dataset, performance of our approach with selected points yielded an accuracy of 95%, specificity of 94.9%, and sensitivity of 95.6% for these classification techniques.

It must be pointed out that at the proteomic level; there may be two types of biomarkers that can be related to cancer. It could be that cancer results in the presence of specific proteins, which are not present in the non-cancer cases. The cancer can be diagnosed by detecting the physical presence of these specific proteins. Alternatively, the cancer may not lead to expression of a novel protein. Instead, it may change the complex proteomic pattern of the tumor-host microenvironment. In this case, the biomarker may be those normal host proteins that are aberrantly increased or decreased in abundance. This is an application where the feature extraction-selection techniques can be most helpful.

REFERENCES


