Feature Extraction and Analysis of Prostate Cancer Proteomic Mass Spectra for Biomarker Discovery

Panagiotis Bougioukos, Dionisis Cavouras, Antonis Daskalakis, Sofia Kossida, George Nikiforidis and Anastasios Bezerianos

***Abstract*— Early detection of cancer is a critical issue for improving patient survival rates. Recent progress in mass spectrometry has shown the promising potential of biomarker discovery in the diagnosis of diseases especially in early stages. In the present study, an alternative approach to feature extraction from mass spectrometry data of prostate cancer is proposed that results in the definition of different biomarkers. The latter provide information rich features that improve the performance of an MLP classifier in differentiating among datasets with different PSA levels of prostate cancer and with no evidence of disease. Prostate cancer dataset was collected from the National Cancer Institute Clinical Proteomics Database. The overall accuracy, in correctly classifying 63 spectra with no evidence of disease (PSA<1) and 69 spectra with prostate cancer (PSA≥4), was 95%. Furthermore 93% was the classification overall accuracy in discriminating 26 spectra of prostate cancer with (4 ≤ PSA<10) from 43 spectra of prostate cancer with (PSA>10). The high accuracies obtained by the proposed method might lead to informative biomarkers for early stage of prostate cancer diagnosis.**

I. Introduction

EARLY detection of cancer is a critical issue for improving patient survival rates. Prostate cancer is a very common cancer disease. The most widely used method for prostate cancer detection is measuring the concentration of the prostate specific antigen (PSA). PSA is the best marker used in clinical practice. The method has the desirable property of yielding high sensitivity but also the drawback that specificity is relatively low. This implies that most of the patients with prostate cancer will be diagnosed correctly but also several non-cancer patients will be diagnosed to have cancer [1].

Recently, several studies have focused on a relatively new technique such as, Surface Enhanced Laser Desorption Ionization (SELDI) mass spectrometry (MS) and Matrix Assisted Laser Desorption Ionization (MALDI) mass spectrometry (MS). The output produced from those techniques concerns mass spectra: x-axis refers to mass to charge ratio (m/z) of proteins and y-axis to their relative intensity. Processing of these data is an alternative approach to prostate cancer detection. Mass spectrometry enables rapid identification of differentially expressed or altered proteins. Recent progress in mass spectrometry has shown the promising potential of biomarker discovery in the diagnosis of diseases especially in early stages [2]-[19].

MS data analysis concerns the ability to detect biomarkers. A biomarker is an identified protein, which is correlated with the state of a particular disease or condition [18]. A common two step approach focuses on spectral peaks. The first step involves feature extraction and quantification, in which one identifies the peak locations and quantifies each peak. This requires one to deal with several modeling issues, including calibration of the spectra, baseline correction, smoothing and normalization. An important issue with processing mass spectra is how to handle data with the number of variables (the different protein masses on the x-axis) being much greater than the number of available samples. This calls for the use of dimensionality reduction techniques, which must be carried out before pattern recognition (or classification) algorithms can be applied on the data. The second step consists of using the generated feature dataset in order to apply unsupervised clustering or supervised learning methods to perform discrimination and classification.

Previous studies have focused on the classification [2]-[4], [6], [9]-[14], [17], [19] part (separating cancer samples from healthy) and others have put more effort in finding specific important proteins or peptides [2], [7], [8], [14], [17], [19]. In [4], the SELDI software program has been employed to detect spectral peaks and a two-sided Wilcoxon test was used to discriminate healthy individuals from patients with ovarian cancer. Coombes et al. [8] have used a method including principal component analysis (PCA) to analyze a sample set from breast cancer patients. Ball et al. [10] have used SELDI together with Neural Nets to examine and classify different types of brain tumors. Rogers et.al. [12] have used a peak detection algorithm and a feed-forward neural network classifier to distinguish renal cancer from healthy renal tissue. Wang et al [13] have applied a feature extraction technique in head and neck cancer dataset, based on finding peaks, and have employed a classification scheme consisting of five classifiers: logistic regression, majority k-nearest neighbor, generalized regression neural network, multi-layer perceptron classifier neural network,

Manuscript received June 30, 2006. This work was supported in part by the General Secretariat for Research and Technology, Greece (Grant PENED 2003/136 jointly funded with the European Union).

P. Bougioukos, A. Daskalakis, George Nikiforidis and A.Bezerianos are with the , Department of Medical Physics, School of Medicine, University of Patras, Rio , GR-26503 Greece (correspondence author; phone: 2610-997745; e-mail: bougiouk@upatras.gr).

Dionisis Cavouras is with the Medical Signal and Image Processing Lab, Department of Medical Instrumentation Technology,Technological Education Institution of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece (e-mail: cavouras@teiath.gr).

Sofia Kossida is with the Division of Biotechnology, Foundation for Biomedical Research, Academy of Athens, Athens, Greece.

and linear support vector machine. Each classifier provided a score for a test subject, which is the probability that the given subject has the disease. The median of the five scores for each subject was used to classify the subject as cancerous or normal.

Regarding prostate cancer, Yasui et al. [7] have proposed peak detection followed by a boosting algorithm to analyze prostate cancer samples and controls. Petricoin et al. have produced data sets (spectra) from both prostate and ovarian cancer patients and have used cluster analysis, combined with genetic algorithms, to classify the samples [2], [3]. Adam et al [19] have used the Ciphergen SELDI software to detect peaks and a decision tree algorithm to discriminate prostate cancer from healthy patterns. In another study [5], the Discrete Wavelet Transform has been employed for data dimensionality reduction and the Fisher Linear Discriminant (FLD) for distinguishing healthy from patients with prostate cancer. In [6], the Boosted Decision Stump Feature Selection and the AdaBoost classifiers have been applied to discriminate patients with prostate cancer from healthy. Lilien et al. [11] have developed a method called Q5 that includes dimensionality reduction with PCA and the FLD classifier to distinguish ovarian cancer from controls and prostate cancer from controls. In these studies, several biomarkers have been proposed, which however, differ.

In the present study, an alternative approach to feature extraction from mass spectrometry data of prostate cancer is proposed that results in the definition of different biomarkers. The latter provide information rich features that improve the performance of an MLP classifier in differentiating among datasets with different PSA levels of prostate cancer and with no evidence of disease.

II. Materials and Methods

Prostate cancer dataset was collected from the National Cancer Institute Clinical Proteomics Database. Data were produced using the H4 protein chip and a Ciphergen PBS1 SELDI-TOF mass spectrometer. The chip was prepared by hand and spectra were exported with baseline subtracted. Collected dataset comprised 190 serum spectra from patients with benign prostate (PSA > 4), 63 spectra with no evidence of disease (PSA<1), 26 spectra with prostate cancer ($4 \leq$ PSA<10) and 43 spectra with prostate cancer (PSA>10). Each spectrum is a histogram with 15.156 m/z data points with each data point corresponding to a single feature.

A. Smoothing

Signal noise contamination was reduced by the lowess smoothing technique [20]. Accordingly, a sliding window with size equal to 1% of the number of points in the x-axis (m/z) values was utilized over the processed signal in order to perform weighted linear squares fit on the points contained in the current window.

B. Thresholding

Spectral intensity values were suitably thresholded for keeping the most significant values in the spectrum. The histogram of each spectrum was calculated and the threshold's maximum value, depicting the average mass spectrum intensity level, determined the threshold, below which all intensity values in the spectrum were zeroed. Thus, by introducing a threshold, the number of intensity values in each mass spectrum was reduced from 15.154 data points to approximately to 5.000-7.000 data points [13].

C. Feature Extraction

A peak detection technique was applied, based on searching for local maxima (features) among the modified spectra, applying a differentiation method between successive intensity data points. Thus, the number of significant intensity values (features) was reduced to 60-80 data points for each spectrum (see Figure.1). The varying number of peaks was due to chemical and electronic noise [16]. To alleviate this, a peak alignment process was developed, that aligned peaks appearing concurrently in all the available spectra, but sustaining a small shift along the x-axis, and ignored the rest. At the end, an equal number of aligned peaks appeared in each mass spectrum.

Accordingly,

1/the local maxima (peaks) of each mass spectrum formed the spectrum's feature vector. The vector with the smallest number of peaks was set to be the *reference* vector.

2/by scanning all feature vectors, the smallest distance $\boldsymbol{d_{min}}$ between two successive peaks was determined.

3/ $\boldsymbol{d_{min}}$ was used to form a 2* $\boldsymbol{d_{min}}$ interval centered at the first peak of the reference vector. Within that interval all feature vectors were scanned and if over than 50% of the vectors contained peaks, then that peak value of the reference vector was considered as a significant feature (biomarker).

4/step 3 was performed for all peaks of the reference vector, thus, providing a number of biomarkers, which were used as significant features to represent each mass spectrum. When a thus chosen biomarker was not present in a feature vector, then its value was calculated as the mean of the corresponding non-zero intensities.

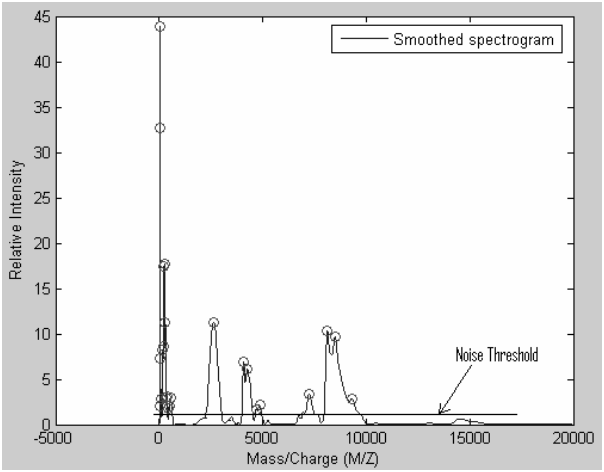

Fig.1. Local maxima (peaks) are specified with the circle symbol and the estimated noise threshold.

D. Classification

Classification was performed by means of a multilayer perceptron classifier (MLP), which was trained to discriminate spectra with no evidence of disease (PSA<1) and spectra with prostate cancer (PSA≥4). Additionally, the MLP was trained to distinguish between spectra of prostate cancer with PSA $4 \leq$ PSA<10 and PSA>10. The MLP structure was 5-5-1, with input the feature vector from each mass spectrum. The classification accuracy of the MLP was evaluated by means of the leave-one-out method where each time a vector was left-out, the MLP was designed with the rest of the vectors, and the left-out was classified by the MLP. This procedure was repeated for all vectors and the classification results were presented in a truth table.

III. Results

The overall accuracy, in correctly classifying 63 spectra with no evidence of disease (PSA<1) and 69 spectra with prostate cancer (PSA≥4), was 95% (see Table I).

TABLE I
CLASSIFICATION RESULTS FOR 63 SPECTRA WITH NO EVIDENCE OF DISEASE (PSA<1) AND 69 SPECTRA WITH PROSTATE CANCER (PSA≥4)

No. Spectra / PSA level	PSA<1	PSA≥4	accuracy
63 / PSA<1	61	2	96%
69 / PSA≥4	4	65	94%
Overall accuracy			95%

$$Sensitivity = \frac{TP}{FN + TP} = 94\%$$

$$Specificity = \frac{TN}{FP + TN} = 97\%$$

Table II shows the overall accuracy (93%) in accurately discriminating 26 spectra of prostate cancer with ($4 \leq$ PSA<10) from 43 spectra of prostate cancer with (PSA>10).

TABLE II
CLASSIFICATION RESULTS FOR 26 SPECTRA OF PROSTATE CANCER WITH ELEVATED PSA LEVEL AGAINST 43 SPECTRA OF PROSTATE CANCER WITH PSA LEVEL GREATER THAN TEN

No. Spectra / PSA level	PSA<1	PSA≥4	accuracy
63 / PSA<1	23	3	88%
69 / PSA≥4	2	41	95%
Overall accuracy			93%

IV. Discussion

The approach followed reduced the dimensionality of mass spectrometry data and determined biomarkers corresponding to proteins that discriminated with high accuracy normal from prostate cancer spectral data. Approximately 10 biomarkers (peaks) amongst 15.154 data points/markers were shown to have high discriminatory ability.

Pre-processing of mass spectrometry data is a very critical step in the overall analysis of MS data set. Smoothing, noise estimation, as well as spectrum alignment and peak detection all affect the performance of classification. Previous studies [3], [6], [19] have also attempted to determine potential biomarkers for prostate cancer disease diagnosis. However, as summarized in Table III, potential biomarkers vary significantly among studies. Considering that they have used the same database, differences may be attributed to variations in pre-processing methods adopted in those studies. The proposed by the present study pre-processing steps (smoothing-peak detection-peak alignment) have revealed mostly different biomarkers, however, one biomarker (4077 m/z) seems to approximate one biomarker obtained by another study (4071 m/z). The ultimate goal is to find the m/z locations in the MS data where the control cases and the disease cases show the most significant differences. The potential biomarkers that are proposed in this study are depicted in Table III.

TABLE III
FEATURES EXTRACTED FROM THE PROSTATE CANCER DATA SET BY THE PROPOSED METHOD. COLUMN 1 SHOWS THE NUMBER OF SIGNIFICANT FEATURES. COLUMN 2 PROVIDES THE FEATURES EXTRACTED FROM THE PRESENT WORK. THE LAST THREE COLUMNS PRESENT FEATURES FOUND IN PREVIOUS STUDIES [3, 6, 19].

Number of features	M/Z	Adam et al.	Qu et al.	Petricoin et al.
1	28	4475	3486	3080
2	47	5074	3963	4819
3	96		4071	5439
4	194		4080	
5	234		5289	
6	252			
7	299			
8	342			
9	368			
10	4077			

The accuracies obtained by the proposed method presented in this study, demonstrate that SELDI protein chip mass spectrometry combined with an MLP neural network classification algorithm can both facilitate the determination of informative biomarkers for prostate cancer providing an innovative clinical diagnostic platform.

V. Acknowledgements

This work was supported by a grant form the General Secretariat for Research and Technology, Ministry of Development of Greece (013/PENED03) to A.B.

References

[1] L. L. Banez, S. Srivastava, and J. W. Moul, "Proteomics in prostate cancer," *Curr Opin Urol*, vol. 15, pp. 151-6, 2005.

[2] E. F. Petricoin, A. M. Ardekani, B. A. Hitt, P. J. Levine, V. A. Fusaro, S. M. Steinberg, G. B. Mills, C. Simone, D. A. Fishman, E. C. Kohn, and L. A. Liotta, "Use of proteomic patterns in serum to identify ovarian cancer," *Lancet*, vol. 359, pp. 572-7, 2002.

[3] E. F. Petricoin, 3rd, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velassco, C. Trucco, L. Wiegand, K. Wood, C. B. Simone, P. J. Levine, W. M. Linehan, M. R. Emmert-Buck, S. M. Steinberg, E. C. Kohn, and L. A. Liotta, "Serum proteomic patterns for detection of prostate cancer," *J Natl Cancer Inst*, vol. 94, pp. 1576-8, 2002.

[4] J. M. Sorace and M. Zhan, "A data review and re-assessment of ovarian cancer serum proteomic profiling," *BMC Bioinformatics*, vol. 4, pp. 24, 2003.

[5] Y. Qu, B. L. Adam, M. Thornquist, J. D. Potter, M. L. Thompson, Y. Yasui, J. Davis, P. F. Schellhammer, L. Cazares, M. Clements, G. L. Wright, Jr., and Z. Feng, "Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data," *Biometrics*, vol. 59, pp. 143-51, 2003.

[6] Y. Qu, B. L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z. Feng, O. J. Semmes, and G. L. Wright, Jr., "Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients," *Clin Chem*, vol. 48, pp. 1835-43, 2002.

[7] Y. Yasui, M. Pepe, M. L. Thompson, B. L. Adam, G. L. Wright, Jr., Y. Qu, J. D. Potter, M. Winget, M. Thornquist, and Z. Feng, "A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection," *Biostatistics*, vol. 4, pp. 449-63, 2003.

[8] K. R. Coombes, H. A. Fritsche, Jr., C. Clarke, J. N. Chen, K. A. Baggerly, J. S. Morris, L. C. Xiao, M. C. Hung, and H. M. Kuerer, "Quality control and peak finding for proteomics data collected from nipple aspirate fluid by surface-enhanced laser desorption and ionization," *Clin Chem*, vol. 49, pp. 1615-23, 2003.

[9] J. H. Oh, J. Gao, A. Nandi, P. Gurnani, L. Knowles, J. Schorge, and K. P. Rosenblatt, "Diagnosis of Early Relapse in Ovarian Cancer Using Serum Proteomic Profiling," *Genome Informatics*, vol. 16, pp. 195-204, 2005.

[10] G. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, and R. C. Rees, "An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers," *Bioinformatics*, vol. 18, pp. 395-404, 2002.

[11] R. H. Lilien, H. Farid, and B. R. Donald, "Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum," *J Comput Biol*, vol. 10, pp. 925-46, 2003.

[12] M. A. Rogers, P. Clarke, J. Noble, N. P. Munro, A. Paul, P. J. Selby, and R. E. Banks, "Proteomic profiling of urinary proteins in renal cancer by surface enhanced laser desorption ionization and neural-network analysis: identification of key issues affecting potential clinical utility," *Cancer Res*, vol. 63, pp. 6971-83, 2003.

[13] X. Wang, W. Zhu, K. Pradhan, C. Ji, Y. Ma, O. J. Semmes, J. Glimm, and J. Mitchell, "Feature extraction in the analysis of proteomic mass spectra," *Proteomics*, vol. 6, pp. 2095-100, 2006.

[14] K. A. Baggerly, J. S. Morris, J. Wang, D. Gold, L. C. Xiao, and K. R. Coombes, "A comprehensive approach to the analysis of matrix-assisted laser desorption/ionization-time of flight proteomics spectra from serum samples," *Proteomics*, vol. 3, pp. 1667-72, 2003.

[15] K. A. Baggerly, J. S. Morris, and K. R. Coombes, "Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments," *Bioinformatics*, vol. 20, pp. 777-85, 2004.

[16] M. Hilario, A. Kalousis, C. Pellegrini, and M. Muller, "Processing and classification of protein mass spectra," *Mass Spectrom Rev*, vol. 25, pp. 409-49, 2006.

[17] W. Yu, B. Wu, N. Lin, K. Stone, K. Williams, and H. Zhao, "Detecting and aligning peaks in mass spectrometry data with applications to MALDI," *Comput Biol Chem*, vol. 30, pp. 27-38, 2006.

[18] J. S. Morris, K. R. Coombes, J. Koomen, K. A. Baggerly, and R. Kobayashi, "Feature extraction and quantification for mass spectrometry in biomedical applications using the mean spectrum," *Bioinformatics*, vol. 21, pp. 1764-75, 2005.

[19] B. L. Adam, Y. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z. Feng, and G. L. Wright, Jr., "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men," *Cancer Res*, vol. 62, pp. 3609-14, 2002.

[20] www.mathworks.com.