

Biomarker Selection, Employing an Iterative Peak Selection Method, and Prostate Spectra Characterization for Identifying Biomarkers Related to Prostate Cancer

Panagiotis Bougioukos¹, Dionisis Cavouras², Antonis Daskalakis¹,
Ioannis Kalatzis², George Nikiforidis¹, and Anastasios Bezerianos¹.

¹ Department of Medical Physics, School of Medicine, University of Patras, Rio ,
GR-26500 Greece

² Medical Signal and Image Processing Lab, Department of Medical Instruments Technology,
Technological Educational Institute of Athens, Ag. Spyridonos Street, Aigaleo,
122 10, Athens, Greece

Abstract. A proteomic analysis system (PAS) for prostate Mass Spectrometry (MS) spectra is proposed for differentiating normal from abnormal and benign from malignant cases and for identifying biomarkers related to prostate cancer. PAS comprised two stages, 1/a pre-processing stage, consisting of MS-spectrum smoothing, normalization, iterative peak selection, and peak alignment, and 2/a classification stage, comprising a 2-level hierarchical tree structure, employing the PNN and SVM classifiers at the 1st (normal-abnormal) and 2nd (benign-malignant) classification levels respectively. PAS first applied local thresholding, for determining the MS-spectrum noise level, and second an iterative global threshold estimation algorithm, for selecting peaks at different intensity ranges. Two optimum sub-sets of these peaks, one at each global threshold, were used to optimally design the hierarchical classification scheme and, thus, indicate the best m/z values. The information rich biomarkers 1160.8, 2082.2, 3595.9, 4275.3, 5817.3, 7653.2, that have been associated with the prostate gland, are proposed for further investigation.

Keywords: Mass-Spectrometry; Biomarker Selection; Classification.

1 Introduction

Early detection of cancer is a critical issue for improving patient survival rates. Regarding prostate cancer, it is the second leading cause of cancer deaths in United States and Canada [1]. In daily clinical practice, the most widely used method for prostate cancer detection is the measurement of the prostate specific antigen (PSA). The PSA diagnostic test exhibits high sensitivity. However, its low specificity confines its use as an early detection biomarker. This calls for the discovery of novel biomarkers that will result in higher specificity, thus, aiding in the decrease of prostate cancer mortality.

With recent advances in mass spectrometry techniques (MS), it is now possible to investigate proteins over a wide range of molecular weights in small biological specimens, such as serum [2]. However, MS proteomic profiles have several imperfections, which

may complicate their interpretation [3]. Thus, data preprocessing steps, such as smoothing, normalization, peak detection, and peak alignment may improve the performance of proteomic analysis and assist in biomarker discovery [4].

Regarding prostate cancer, previous studies [5-10] have implemented various pre-processing algorithms and have applied different pattern recognition techniques for proposing biomarkers, which however differ. Classification accuracies that have been achieved in some of those studies ranged between 73% and 100% [5, 9-11], however, using own data, while in two studies [6, 7], use has been made of the same with us dataset, achieving 97% and 92% classification accuracies. The latter have been attained by the inclusion of less than 1000 m/z values, which, however, have been reported to be of non-significance due to distortion [10].

In the present study, a proteomic analysis system (PAS) for prostate Mass Spectrometry (MS) spectra is proposed for differentiating normal from abnormal and benign from malignant cases and for identifying biomarkers related to prostate cancer. PAS comprised two stages, 1/a pre-processing stage, consisting of MS-spectra smoothing, normalization, iterative peak selection, and peak alignment, and 2/a classification stage, comprising a 2-level hierarchical tree structure employing the PNN and SVM classifiers at the 1st (normal-abnormal) and 2nd (benign-malignant) classification levels respectively. PAS first applied local thresholding, for determining the MS-spectrum noise level, and second an iterative global threshold estimation algorithm, for selecting peaks at different intensity ranges. Two optimum sub-sets of these peaks, one at each global threshold, were used to optimally design the hierarchical classification scheme. Among those peaks, prostate related biomarkers were identified by reference to published literature and public databases.

2 Materials and Methods

Mass spectrometry prostate cancer spectra were obtained from the National Cancer Institute Clinical Proteomics Database [12]. MS-spectra were produced utilizing the H4 protein chip and a Ciphergen PBS1 SELDI-TOF mass spectrometer. The chip was prepared by hand and MS-spectra were exported with baseline subtracted. The dataset comprised 63 spectra, with no evidence of disease ($PSA < 1$), 69 prostate cancer spectra ($PSA > 4$), and 69 benign spectra ($PSA > 4$).

2.1 PAS Pre-processing Stage

Signal noise residues were diminished applying the Lowess smoothing technique [13] (see Fig. 1), a function provided by Matlab's bioinformatics toolbox. After smoothing, all spectra were normalized to their total ion current and spectra were scaled to have an overall maximum intensity of 100.

Following smoothing, all MS-spectra were further processed utilizing an iterative peak selection method. Accordingly, the steps employed were:

a/ For each spectrum the noise level was estimated by dividing the spectrum into four equal parts. In this way, local thresholds were determined and were further used to determine intensities (above local thresholds) and noise (below local thresholds) (see Fig. 2).

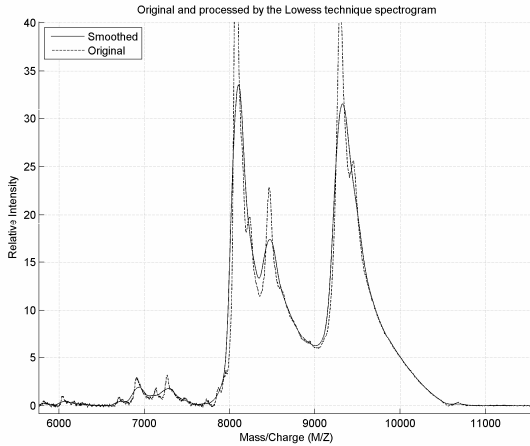


Fig. 1. Original and processed by the Lowess technique spectrogram

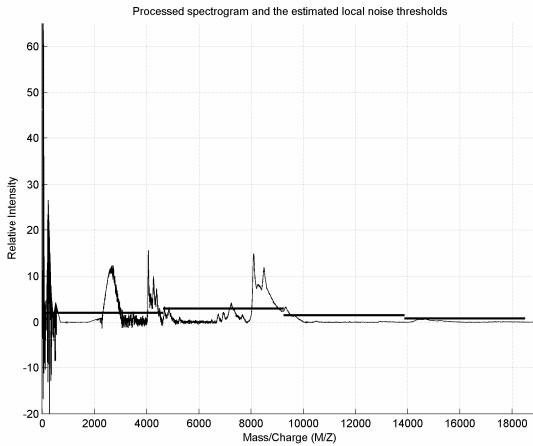


Fig. 2. Local noise threshold estimation for each mass spectrum

Local thresholds were determined by first computing at each quartile the histogram of intensities and then by locating the histogram’s maximum value (see Fig 3).

b/ Intensity values above local thresholds were searched for peaks, by applying a differentiation method between successive intensity data points.

c/ Selected peaks were used to form a histogram of intensities, and its maximum value constituted the global threshold, above which all peaks were considered as most informative and were stored for further processing.

d/ Peaks determined in step **c** were removed from the spectrum and step **c** was repeated until all peaks of the entire spectrum were removed.

Steps **a-d** were repeated for all the MS-spectra of each class. In this way, peaks were grouped according to their global threshold level.

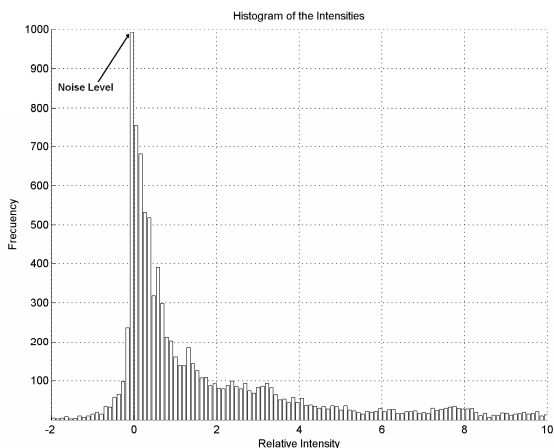


Fig. 3. Histogram of the intensities across m/z values. The maximum of the histogram depicts the average mass spectrum noise intensity level.

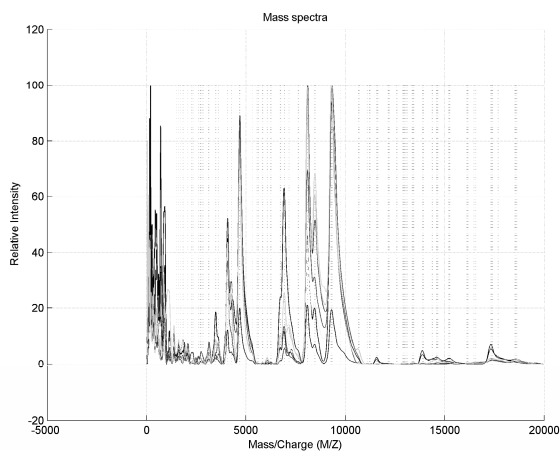


Fig. 4. Seven control cases mass spectra. The dotted vertical lines represent the corresponding m/z values of the peaks.

At each global threshold level, each MS-spectrum was represented by a varying number of peaks due to chemical and electronic noise [14]. To alleviate this, a peak alignment process was developed, that aligned peaks appearing concurrently in 5% of the available spectra but sustaining a small shift along the x-axis, 0.08 % and ignoring the rest (see Fig 4).

After peak alignment at each global threshold, peaks were further used to optimally design PNN and SVM classifiers (see section 2.2).

2.2 PAS Classification Stage

A two-level hierarchical tree structure was constructed for differentiating normal from abnormal (1st level) and benign from malignant (2nd level) cases employing:

1/ At the 1st level a Probabilistic Neural Network (PNN) classifier with discriminant function given by [15] (eq. 1) :

$$d_i(x) = \frac{1}{(2\pi)^{d/2} \sigma^d} \frac{1}{N} \sum_{k=1}^N \exp \left[-\frac{(x - x_{ik})^T (x - x_{ik})}{2\sigma^2} \right] \tag{1}$$

where σ is the spread of the Gaussian activation function that was experimentally determined and was set to 0.5, N is the number of pattern vectors, d is the dimensionality of pattern vectors and x_{ik} is the k^{th} pattern vector of class i .

2/ At the 2nd level, a Support Vector Machines (SVM) classifier was employed with discriminant function [16] given by (eq. 2).

$$g(x) = \text{sign} \left(\sum_{i=1}^N a_i y_i K(x, x_i) + b \right) \tag{2}$$

where x_i is the training data belonging to class i , y_i is the class label $\{-1, +1\}$, N is the number of pattern vectors, a_i , b are weight coefficients and K the transformation or kernel function [16]. The Kernel function utilized was the radial basis function (RBF) (eq.3).

The optimization problem of calculating parameters a_i , was solved by using the function quadprog provided by the MATLAB’s optimization toolbox.

$$K_{RBF}(x, x_i) = \exp \left(\frac{-|x - x_i|^2}{2\sigma^2} \right) \tag{3}$$

At each global threshold level, feature selection for both classifiers was performed in order to retain only the feature subset with the highest discriminatory information. For this purpose, the sequential backward feature selection algorithm was employed [17]. Classifiers’ evaluation was performed by the standard procedure of the leave-one-out method (LOO) [17]. The classifiers were designed employing features from all but one MS-spectrum, which was then classified. The process was repeated, each time leaving-out a different MS-spectrum, until all data were processed. In this way, classifiers were evaluated by spectra not involved in their design. The above method has been widely used in many studies, with limited number of data [11, 18, 19]. Among the set of best features employed in the optimal design of the hierarchical tree structure, prostate related biomarkers were identified by reference to published literature and public databases [20].

3 Results

In distinguishing spectra with no evidence of disease ($PSA < 1$) from spectra of benign and malignant nature ($PSA \geq 4$) PAS scored 94.5% and 93% overall accuracy (see Tables 1 and 2) at the 1st and 2nd global threshold levels. The m/z values corresponding to the best features obtained at the 2 classification levels of the hierarchical tree structure and for the two aforementioned global thresholds are illustrated in Table 3. Table 4 illustrates a subset of the m/z values, shown in Table 3, corresponding to verified biomarkers, which were identified by reference to published literature and public databases [20]. These m/z values are proposed as useful biomarkers.

Table 1. PAS classification results at the first global threshold level, for spectra with no evidence of disease and spectra of benign and malignant nature

	Normal	Benign	Malignant	Accuracy (%)
Normal	62	1	0	98.4
Benign	1	64	4	92.8
Malignant	1	4	64	92.8
Overall accuracy				94.5

Table 2. PAS classification results at the second global threshold level, for spectra with no evidence of disease and spectra of benign and malignant nature

	Normal	Benign	Malignant	Accuracy (%)
Normal	62	1	0	98.4
Benign	4	62	3	89.9
Malignant	3	3	63	91.3
Overall accuracy				93

Table 3. m/z values of the best feature vectors

Global Threshold levels	Feature vectors (biomarkers)
First	{ 1160.8, 1838.2, 3595.9, 4074.3, 4257.1, 4275.3, 6918.2, 8449.1, 8470.2, 9317.5, 9767.5, 14609.9, 4087.9, }
Second	{ 1060.5, 1298.8, 1468.5, 1755.6, 2082.2, 2213.6, 2736.1, 4503.7, 4672, 5326.0, 5817.3, 10673.1, 3607.1, 6939.1, 7653.2 }

Table 4. Verified biomarkers selected from the two best feature vectors

Global Threshold levels	Feature vectors (biomarkers)
First	{ 1160.8, 3595.9, 4275.3 }
Second	{ 2082.2, 4672, 5817.3, 7653.2 }

4 Discussion

The goal of this study was to develop a proteomic analysis system for prostate MS-spectra characterization into normal, benign, and malignant and to propose biomarkers related to prostate cancer. The proposed PAS-system achieved 94.5% and 93% at the first and the second global threshold levels respectively, in discriminating individuals with no evidence of disease ($PSA < 1$) from those with prostate disease (benign or malignant, $PSA \geq 4$). Employing the proposed iterative peak selection method and by investigating the accuracies attained, it was plausible to conclude that peaks realizing high intensity values are not necessarily the most significant in distinguishing normal from abnormal prostate cases. This iterative peak selection method facilitates in reducing the number of potential biomarkers as well as in assisting researchers by focusing on specific intensity levels for biomarker discovery. Specifically, the highest discriminating power was achieved using the biomarkers that were revealed at specific global threshold levels.

Seeking for proteins that might be related to prostate cancer, it was found, by searching in the ExPASy database [20], that the m/z value of 2082.2 was very close to the Nociceptin protein (2081.39 Daltons), which has been implicated in the stimulation of prostate cell growth [21], among other neuropeptides. Additionally, the 4672 m/z value is very close to the BAX protein, cytoplasmic isoform gamma that has a molecular weight of 4678.22 (Daltons). This protein belongs to the BCL-2 gene family; however, as it has been reported in [22], elevated levels of BCL-2 might aid in the progression of the prostate cancer. Also, another m/z value (biomarker) was the 5817.3, which is very close to the 5818.62 (Granulin precursor). This precursor is a prostate cancer cell-derived growth factor and its expression has been found to be elevated in high-grade prostatic intraepithelial neoplasia and prostatic adenocarcinoma [23]. Additionally, the biomarker 7653.2 has m/z value close to 7654.74, which is the Insulin-like Growth factor I (IGF-I). This protein is a mitogen for prostate epithelial cells and it has been reported [24] to have a strong positive association between elevated levels of IGF-I and prostate cancer risk. The 1160.8 m/z value is close to 1169.32 biological marker, named prostasin precursor, which is present in many tissues and has the highest level in prostate gland [25]. One of the seminal plasma proteins realizing 3590.98 m/z value is very close to the 3595.9 m/z value, which was identified to be seminal basic protein (BSP), a proteolytic product of semenogelin 1 [26]. The 4275.3 has m/z value close to 4272.2, which is the Neuropeptide Y precursor and its activation might be related in cell proliferation in certain stages of prostate cancer [27]. Neuropeptide Y precursor (NPY) (3474.91 daltons) is close to 3474 m/z value and it has been reported [28] that NPY may directly regulate prostate cancer growth via Y1-R gene expression. This finding suggests that NPY related mechanism might play a relevant role in the progression of prostate cancer at both androgen dependent and independent stages. The present study revealed other m/z values as being significant in distinguishing prostate cancer dataset, as illustrated in Tables 3. These values might constitute information rich biomarkers that have not been yet identified or related to prostate cancer.

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

References

1. McDavid, K., Lee, J., Fulton, J.P., Tonita, J., Thompson, T.D.: Prostate cancer incidence and mortality rates and trends in the united states and canada. *Public Health Rep.* 119, 174–186 (2004)
2. Srinivas, P.R., Srivastava, S., Hanash, S., Wright Jr, G.L.: Proteomics in early detection of cancer. *Clin Chem.* 47, 1901–1911 (2001)
3. Hilario, M., Kalousis, A., Pellegrini, C., Muller, M.: Processing and classification of protein mass spectra. *Mass Spectrom Rev.* 25, 409–449 (2006)
4. Malyarenko, D.I., Cooke, W.E., Adam, B.L., Malik, G., Chen, H., Tracy, E.R., Trosset, M.W., Sasinowski, M., Semmes, O.J., Manos, D.M.: Enhancement of sensitivity and resolution of surface-enhanced laser desorption/ionization time-of-flight mass spectrometric records for serum peptides using time-series analysis techniques. *Clin Chem.* 51, 65–74 (2005)
5. Jong, K., Marchiori, E., Sebag, M., van der Vaart, A.: Feature selection in proteomic pattern data with support vector machines. In: *Proceedings of the* (2004) 41
6. Lilien, R.H., Farid, H., Donald, B.R.: Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum. *J Comput Biol.* 10, 925–946 (2003)
7. Petricoin, E.F., Ornstein 3rd, D.K., Paweletz, C.P., Ardekani, A., Hackett, P.S., Hitt, B.A., Velasco, 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, 1576–1578 (2002)
8. Qu, Y., Adam, B.L., Thornquist, M., Potter, J.D., Thompson, M.L., Yasui, Y., Davis, J., Schellhammer, P.F., Cazares, L., Clements, M., Wright Jr, G.L., Feng, Z.: Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* 59, 143–151 (2003)
9. Qu, Y., Adam, B.L., Yasui, Y., Ward, M.D., Cazares, L.H., Schellhammer, P.F., Feng, Z., Semmes, O.J., Wright Jr, G.L.: Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from non-cancer patients. *Clin Chem.* 48, 1835–1843 (2002)
10. Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright, G.L., Jr., Qu, Y., Potter, J.D., Winget, M., Thornquist, M., Feng, Z.: A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection. *Biostatistics* 4, 449–463 (2003)
11. Adam, B.L., Qu, Y., Davis, J.W., Ward, M.D., Clements, M.A., Cazares, L.H., Semmes, O.J., Schellhammer, P.F., Yasui, Y., Feng, Z., Wright Jr, G.L.: Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res.* 62, 3609–3614 (2002)
12. Institute, N.C. (Accessed 24/11/2006) via the INTERNET Available: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>
13. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. *J Amer Statist Assoc.* 74, 829–836 (1979)

14. Baggerly, K.A., Morris, J.S., Coombes, K.R.: Reproducibility of seldi-tof protein patterns in serum: Comparing datasets from different experiments. *Bioinformatics* 20, 777–785 (2004)
15. Specht, D.F.: Probabilistic neural networks. *Neural Networks* 3, 109–118 (1990)
16. Christianini, N., Taylor, J.S.: An introduction to support vector machines and other kernel-based learning methods. Cambridge University Press, Cambridge (2000)
17. Theodorides, S., Koutroumbas, K.: Pattern recognition, 2nd edn. Academic Press, London (2003)
18. Resson, H.W., Varghese, R.S., Abdel-Hamid, M., Eissa, S.A., Saha, D., Goldman, L., Petricoin, E.F., Conrads, T.P., Veenstra, T.D., Loffredo, C.A., Goldman, R.: Analysis of mass spectral serum profiles for biomarker selection. *Bioinformatics*. 21, 4039–4045 (2005)
19. Wang, X., Zhu, W., Pradhan, K., Ji, C., Ma, Y., Semmes, O.J., Glimm, J., Mitchell, J.: Feature extraction in the analysis of proteomic mass spectra. *Proteomics* 6, 2095–2100 (2006)
20. ExPASy. (Accessed 05/12/2006) Available via the INTERNET: <http://au.expasy.org/tools/>
21. Swanson, T.A., Kim, S.I., Myers, M., Pabon, A., Philibert, K.D., Wang, M., Glucksman, M.J.: The role of neuropeptide processing enzymes in endocrine (prostate) cancer: Ec 3.4.24.15 (ep24.15). *Protein Pept Lett.* 11, 471–478 (2004)
22. Hering, F.L., Lipay, M.V., Lipay, M.A., Rodrigues, P.R., Nesralah, L.J., Srougi, M.: Comparison of positivity frequency of bcl-2 expression in prostate adenocarcinoma with low and high gleason score. *Sao Paulo Med J.* 119, 138–141 (2001)
23. Pan, C.X., Kinch, M.S., Kiener, P.A., Langermann, S., Serrero, G., Sun, L., Corvera, J., Sweeney, C.J., Li, L., Zhang, S., Baldrige, L.A., Jones, T.D., Koch, M.O., Ulbright, T.M., Eble, J.N., Cheng, L.: Pc cell-derived growth factor expression in prostatic intraepithelial neoplasia and prostatic adenocarcinoma. *Clin Cancer Res.* 10, 1333–1337 (2004)
24. Chan, J.M., Stampfer, M.J., Giovannucci, E., Gann, P.H., Ma, J., Wilkinson, P., Hennekens, C.H., Pollak, M.: Plasma insulin-like growth factor-i and prostate cancer risk: A prospective study. *Science* 279, 563–566 (1998)
25. Yu, J.X., Chao, L., Chao, J.: Molecular cloning, tissue-specific expression, and cellular localization of human prostatic mrna. *J Biol Chem.* 270, 13483–13489 (1995)
26. Adam, B.L., Vlahou, A., Semmes, O.J., Wright Jr, G.L.: Proteomic approaches to biomarker discovery in prostate and bladder cancers. *Proteomics* 1, 1264–1270 (2001)
27. Magni, P., Motta, M.: Expression of neuropeptide y receptors in human prostate cancer cells. *Ann Oncol.* 12(2), S27–S29 (2001)
28. Ruscica, M., Dozio, E., Boghossian, S., Bovo, G., Martos Riano, V., Motta, M., Magni, P.: Activation of the y1 receptor by neuropeptide y regulates the growth of prostate cancer cells. *Endocrinology* 147, 1466–1473 (2006)