

Biomarker Selection System, Employing an Iterative Peak Selection Method, for Identifying Biomarkers Related to Prostate Cancer

Panagiotis Bougioukos¹, Dionisis Cavouras², Antonis Daskalakis¹, Ioannis Kalatzis², Spiros Kostopoulos¹, Pantelis Georgiadis¹, 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 Instrumentation Technology, Technological Education Institution of Athens, Ag. Spyridonos Street, Aigaleo, 122 10, Athens, Greece

Abstract. A biomarker selection system is proposed for identifying biomarkers related to prostate cancer. MS-spectra were obtained from the National Cancer Institute Clinical Proteomics Database. The system comprised two stages, a pre-processing stage, which is a sequence of MS-processing steps consisting of MS-spectrum smoothing, novel iterative peak selection, peak alignment, and a classification stage employing the PNN classifier. The proposed iterative peak selection method was based on first applying local thresholding, for determining the MS-spectrum noise level, and second applying an iterative global threshold estimation algorithm, for selecting peaks at different intensity ranges. At each global threshold, an optimum sub-set of these peaks was used to design the PNN classifier for highest performance, in discriminating normal cases from cases with prostate cancer, and thus indicate the best m/z values. Among these 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

Prostate cancer is the second leading cause of cancer in the United States and Canada [1]. In daily clinical practice, prostate specific antigen (PSA) is the most commonly used diagnostic serum biomarker for detecting prostate cancer in men [2]. Although PSA levels can detect prostate cancer, this biomarker exhibits poor specificity, as infections of the prostate gland, such as prostatitis, can also elevate PSA levels in the absence of cancer [3]. Due to a lack of accurate biomarkers that detect, quantify, and reliable distinguish patients with different prostate cancer stages, many early stage prostate cancer patients are treated as having an aggressive state of prostate cancer [4]. This fact necessitates the discovery of new biomarkers that aid in the early detection of prostate cancer.

Mass spectrometry (MS) proteomic profiles enable rapid identification of differentially expressed or altered proteins (biomarkers). A significant problem in MS-spectra is the large number of points containing information rich biomarkers that need to be extracted. The steps that are usually taken for choosing the right biomarkers are: baseline correction, normalization, smoothing, peak detection, and peak alignment [5, 6]. All these steps are necessary since MS data have several imperfections that complicate biomarker identification and subsequent MS interpretation [7].

Regarding prostate cancer, previous studies [8-13] 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% [8, 12-14], however, using own data, while in two studies [9, 11], use has been made of the same with us dataset, achieving 97% and 92% classification accuracies respectively. 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 [13].

In the present study, a biomarker selection system is proposed for identifying biomarkers related to prostate cancer. The system comprises two stages, a pre-processing stage, which is a pipeline of MS-processing steps consisting of MS-profile smoothing, iterative peak selection, and peak alignment, and a classification stage, which proposes the best m/z values of the MS-profile that discriminate between normal and prostate cancer MS-spectra. Among those best m/z values, information rich biomarkers related to prostate cancer are proposed.

The contribution of the present study lies in a/the design of a high precision pattern recognition system for characterizing prostate MS-spectra as normal or cancerous and b/the proposal of a particular combination of prostate cancer related biomarkers that may be of high biological value in early prostate cancer diagnosis, and c/ the suggestion of a new feature (m/z values) extraction technique based on an iterative peak selection method, to feed the Probabilistic Neural Network (PNN) classifier with corresponding, to m/z values, intensities. The proposed method is based on first applying local thresholding, for determining the MS-profile noise level and, second, applying an iterative global threshold estimation algorithm, for selecting peaks at different intensity ranges. At each global threshold, an optimum sub-set of these peaks was used to design the PNN classifier for highest performance, in discriminating normal cases from cases with prostate cancer. Out of these subsets, 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 [15]. 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), and 69 prostate cancer spectra, with PSA values greater than 4.

2.1 Smoothing

Signal noise residues were diminished applying the Lowess smoothing technique [16] a function provided by Matlab's bioinformatics toolbox.

2.2 Iterative Peak Selection Method

After smoothing all MS-spectra, the proposed iterative peak selection method was applied on each spectrum that comprised the following steps:

1. The spectrum was divided into four equal parts in order to find local thresholds, which will be subsequently used in the determination of the MS-profile's noise level, which, in turn, will assist in the accurate detection of peaks [17]. At each quartile, a local threshold level was determined, below which the signal was considered as noise and it was not taken under consideration for further processing. This local threshold was found from the quartile's histogram of intensities and it was set equal to the intensity value with the highest frequency. Figure 1 shows a quartile histogram for determining local threshold level.
2. Intensity values above local thresholds were searched for peaks, by applying a differentiation method between successive intensity data points.
3. 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.
4. Peaks determined in step 3 were removed from the spectrum and Step 3 was repeated until all peaks of the entire spectrum were removed.

Steps 1-4 were repeated for all the MS-spectra of each class. In this way, peaks were grouped according to their global threshold level.

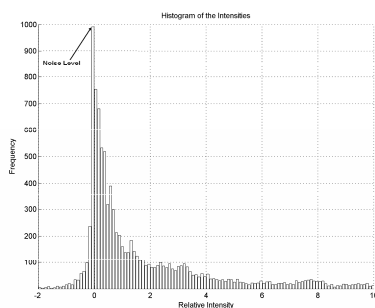


Fig. 1. Histogram of the intensities across m/z values. The maximum of the histogram depicts the average mass spectrum noise intensity level.

2.3 Peak Alignment

Following the feature extraction step, each MS-spectrum was represented by a varying number of peaks due to chemical and electronic noise [18]. To alleviate this, a peak alignment process was developed, that aligned peaks appearing concurrently in

all the available spectra and corresponding to the same global threshold level, but sustaining a small shift along the x-axis, 0.08 % and ignoring the rest. At the end, an equal number of aligned peaks appeared in each MS-spectrum [17] and for the same global threshold level. Peak intensities constituted features that were used as input to the PNN classifier.

2.4 Classification

Classification was performed by means of the Probabilistic Neural Network (PNN) classifier with discriminant function given by [19] (eq. 1) :

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

where σ is the spread of the Gaussian activation function that was experimentally determined to provide highest classification accuracy and was found to be $\sigma = 0.5$, 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 .

The PNN was trained to discriminate controls with no evidence of disease (PSA<1) versus individuals with prostate cancer (PSA>4). At each global threshold level, intensities obtained following the proposed iterative peak selection method were used as input to the PNN classifier. Features were initially normalized to zero mean and unit variance. Subsequently, feature selection 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 [20]. Systems' evaluation was performed by the standard procedure of the leave-one-out method (LOO) [20]. The classifier was 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, the classifier was evaluated by spectra not involved in its design. The above method has been widely used in many studies, with limited number of data [14, 17, 21].

Classification was initially performed at each global threshold level and two sets of best features, providing highest classification accuracy at each level, were extracted. These sets incorporated features, identified as prostate cancer related biomarkers by reference to published literature and public databases [22] as well as uncharacterized peaks. To investigate the impact of biological confirmed markers in prostate cancer, identified biomarkers formed a new input feature-set to train the PNN, once more, for locating the subset of verified biomarkers with the highest discriminatory power.

3 Results

In distinguishing spectra with no evidence of disease (PSA<1) from spectra with prostate cancer (PSA≥4), at the first global threshold level, the PNN classifier scored 96.2% overall accuracy (see Table 1). At the second global threshold, the PNN scored 96.9% overall accuracy (see Table 2). The m/z values corresponding to the best

features, using 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 [22]. A subset among these verified biomarkers was found to provide highest discrimination accuracy with 90.1% (see Table 5) classification accuracy. This set of m/z values {1160.8, 2082.2, 3595.9, 4275.3, 5817.3, 7653.2} are proposed as useful biomarkers for prostate cancer and they should be further investigated.

Table 1. PNN classification results at the first global threshold level, for spectra with no evidence of disease (PSA<1) and spectra with prostate cancer (PSA≥4)

PSA level	PSA<1	PSA≥4	Accuracy
PSA<1	61	2	96.8%
PSA≥4	3	66	95.6%
Overall accuracy			96.2%

Table 2. PNN classification results at the second global threshold level, for spectra with no evidence of disease (PSA<1) and spectra with prostate cancer (PSA≥4)

PSA level	PSA<1	PSA≥4	Accuracy
PSA<1	62	1	98.4%
PSA≥4	3	66	95.6%
Overall accuracy			96.9%

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 }

Table 5. PNN classification employing as features the intensity values of the verified biomarkers {1160.8, 2082.2, 3595.9, 4275.3, 5817.3, 7653.2} that are related to prostate gland

PSA level	PSA<1	PSA≥4	Accuracy
PSA<1	58	5	92.1%
PSA≥4	8	61	88.4%
Overall accuracy			90.1%

4 Discussion

The goal of this study was to propose biomarkers, related to prostate cancer, that discriminate between normal and cancerous cases. Within this context, the contribution of the proposed iterative peak selection method (see section 2.2) was evaluated by means of a probabilistic neural network (PNN) classifier. The PNN classifier achieved 96.2% and 96.9% for the first and the second global threshold levels respectively, in discriminating individuals with no evidence of disease with $PSA < 1$ from those with prostate cancer having $PSA \geq 4$. 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 prostate cancer 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.

Additionally, the system's performance was tested adopting the external cross validation method [23], whereby system's evaluation was performed on test data sets (1/3 of the available data randomly chosen from both classes) that had not participated in any way in the design (by 2/3 of the data) of the classification system. System design was repeated 10 times (each time excluding an arbitrary 1/3 of the data to be used as test set) and a mean and variance of classification accuracy were recorded. Results obtained at the first and second global threshold levels were 90.88 ± 2.31 and 91.52 ± 2.27 respectively.

Previous studies [8, 9] employing the same dataset have also suggested biomarkers, which, however, differ. This may be due to the various pre-processing schemes and classifiers employed. Seeking for proteins that might be related to prostate cancer, it was found, by searching in the ExPASy database [22], 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 [24], 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 [25], 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 [26]. 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 [27] 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 [28]. 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 [29]. 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 [30]. 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

- 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)
- Pannek, J., Partin, A.W.: The role of psa and percent free psa for staging and prognosis prediction in clinically localized prostate cancer. *Semin Urol Oncol.* 16, 100–105 (1998)
- Chan, D.W., Sokoll, L.J.: Prostate-specific antigen: Update 1997. *J. Int. Fed. Clin. Chem.* 9, 120–125 (1997)
- Wright, M.E., Han, D.K., Aebersold, R.: Mass spectrometry-based expression profiling of clinical prostate cancer. *Mol. Cell Proteomics* 4, 545–554 (2005)
- Malyarenko, D.I., Cooke, W.E., Adam, B.L., Malik, G., Chen, H., Tracy, T.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)
- Sauve, A.C., Speed, T.P. Normalization, baseline correction and alignment of high-throughput mass spectrometry. In: *Proceedings of the Data Proceedings Gensips* (2004)
- Hilario, M., Kalousis, A., Pellegrini, C., Muller, M.: Processing and classification of protein mass spectra. *Mass Spectrom Rev.* 25, 409–449 (2006)
- Jong, K., Marchiori, E., Sebag, M., van der Vaart, A.: Feature selection in proteomic pattern data with support vector machines. In: *Proceedings of the*, 41 (2004)
- 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)
- Petricoin, E.F., 3rd, O.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)
- 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, G.L., Jr., F.Z.: Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data. *Biometrics* 59, 143–151 (2003)
- 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 noncancer patients. *Clin. Chem.* 48, 1835–1843 (2002)

13. Yasui, Y., Pepe, M., Thompson, M.L., Adam, B.L., Wright, G.L., Jr., Q.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)
14. 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)
15. Institute, N.C. Available: <http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp>
16. Cleveland, W.S.: Robust locally weighted regression and smoothing scatterplots. *J Amer. Statist. Assoc.* 74, 829–836 (1979)
17. 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)
18. 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)
19. Specht, D.F.: Probabilistic neural networks. *Neural Networks* 3, 109–118 (1990)
20. Theodorides, S., Koutroumbas, K.: Pattern recognition, 2nd edn. Academic Press, London (2003)
21. 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)
22. ExpASY. Accessed 05/12/2006 Available: <http://au.expasy.org/tools/> via the INTERNET
23. Ambrose, C., McLachlan, G.J.: Selection bias in gene extraction on the basis of microarray gene-expression data. *Proc. Natl. Acad. Sci. U.S.A.* 99, 6562–6566 (2002)
24. 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)
25. 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)
26. 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)
27. 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)
28. 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)
29. 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)
30. Magni, P., Motta, M.: Expression of neuropeptide y receptors in human prostate cancer cells. *Ann. Oncol.* 12(2), 27–29 (2001)