2nd International Conference on Experiments/Process/System Modelling/Simulation & Optimization
2nd IC-EpsMsO
Athens, 4-7 July, 2007
© IC-EpsMsO

PROTEOMIC MASS SPECTRA CLASSIFICATION FOR BIOMARKER DISCOVERY IN PROSTATE CANCER, EMPLOYING PATTERN RECOGNITION TECHNIQUES.

Panagiotis Bougioukos[1], Dionisis Cavouras[2], Antonis Daskalakis[1], Spiros Kostopoulos[1], Ioannis Kalatzis[2], George Nikiforidis[1] and Anastasios Bezerianos[1].

[1] Department of Medical Physics, School of Medicine, University of Patras, Rio , GR-26500 Greece. e-mail:
[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.
e-mail: cavouras@teiath.gr, web page: http://medisp.bme.teiath.gr

Keywords: Mass-Spectrometry; Biomarker Selection; Classification.

Abstract. *The purpose of the present study was the proposal of novel biomarkers in prostate cancer by analyzing mass spectrometry profiles. The latter were obtained from the National Cancer Institute Clinical Proteomics Database. The proposed method applied first a pre-processing pipeline of smoothing, automatic noise estimation, peak detection, and peak alignment, for improving the choice of information reach biomarkers and, second, a two level hierarchical tree structure classification scheme, where at each level a PNN classifier was optimally designed. At the first level, normal cases were discriminated by the PNN from cases with prostate cancer of PSA≥4 and, at the second level, distinction was made by the PNN between cancerous cases with 4≤PSA<10 and PSA>10. Maximum classification accuracies were 97.7% and 95.6% respectively. These high accuracies were achieved by a set of information reach biomarkers, which included the 2068.8m/z, 4675.6 m/z, and 5824.5 m/z values that have been associated with prostate cancer.[1]*

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 [2]. 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.

Regarding prostate cancer, previous studies [3-8] have implemented various pre-processing algorithms and have applied different classification techniques for proposing biomarkers, which however differ. Classification accuracies achieved were high ranging between 73% and 100%. However, [3, 6-8] have used own data and [4, 5] have used the same with us dataset. The latter achieved 97% and 92% classification accuracies respectively in distinguishing normal from prostate cancer cases, but they have included biomarkers of less than 1000 m/z value, which have been reported to be on non significance due to distortion [8].

The contribution of the present study lies in proposing biomarkers related to prostate cancer that discriminate between normal and cancerous cases and further between cases of elevated but different PSA values. The proposed method applied a pre-processing pipeline of smoothing, automatic noise estimation, peak detection, and peak alignment, for improving the choice of information reach biomarkers (feature extraction), used in the design of the PNN classifier. Prostate classification was then considered as a two level hierarchical tree structure, where at each level a PNN classifier was optimally designed. At the first level, normal cases were discriminated by the PNN from cases with prostate cancer of PSA≥4 and, at the second level, distinction was made by the PNN between cancerous cases with 4≤PSA<10 and PSA>10. Employing the sequential backward feature selection method, biomarkers were further reduced to reveal the optimum combination of biomarkers that gave the highest PNN classification accuracy. Thus, biomarkers related to prostate cancer were proposed at each level of the decision tree

2 MATERIALS AND METHODS

Mass spectrometry prostate cancer profiles were obtained from the National Cancer Institute Clinical Proteomics Database [9]. MS patterns 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 of 63 spectra with no evidence of disease (PSA<1), 26 prostate cancer spectra with PSA values between 4 and 10 and 43 prostate cancer spectra with PSA greater than 10. Subsequently, all spectra were pre-processed in order to eliminate their imperfections (chemical, electronic noise) [10], performing smoothing and noise estimation.

2.1 Smoothing

Signal noise residues were diminished applying the Lowess smoothing technique [11] (see Fig 1.). Accordingly, at each point in the data set, a low-degree polynomial was fit to a subset of the data, using weighted least squares (see eq.1) giving more weight to points near the point whose response was estimated and less weight to points further away. The sum of the residuals *S* was minimized and was expressed by:

$$S = \sum_{i=1}^{n} w_i (y_i - f(x_i))^2 \tag{1}$$

where wi are the weights, yi are the spectra values at xi positions and f(·) is the polynomial function.

Weights were expressed as functions of the variance σ, giving points with a lower variance a greater statistical weight $w_i = 1/\sigma^2_i$. The value of the regression function for the point was then obtained by evaluating the local polynomial using the estimated variables for that data point. The Lowess fit was completed after the regression function values had been computed for each of the n data points.

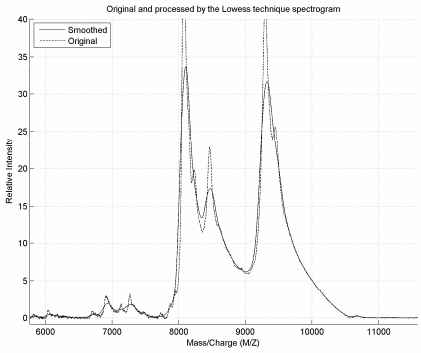

Figure 1. The original spectrogram and the processed one smoothed with the lowess technique.

2.2 Global-Local Noise Estimation

Regarding global noise estimation intensity values were suitably thresholded for keeping the most significant values in the spectrum. The histogram of each spectrum (see Fig 2) 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 not taken into account for further processing (see Fig 3) [12]. Considering local noise estimation for each spectrum, a sliding window was utilized. Window width was selected to be 25% of the whole spectrum, which is the optimum setting to capture true peaks [12]. The histogram of this window was calculated and its maximum value specified the average for the current window intensity level, designating the threshold, below which all intensity values in the spectrum were considered as noise (see Fig 4). The procedure was carried on for each spectrum and across all spectra. Thus, by introducing noise threshold values, the number of intensity values in each mass spectrum was reduced [12].

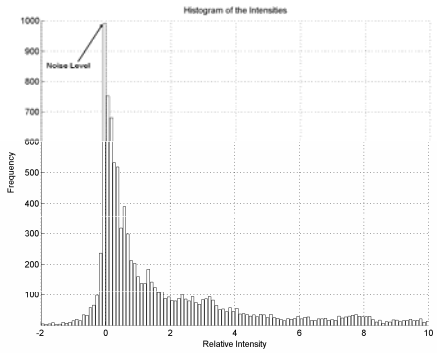

Figure 2. Histogram of the intensities across m/z values. The maximum of the histogram depicts the average mass spectrum noise intensity level.

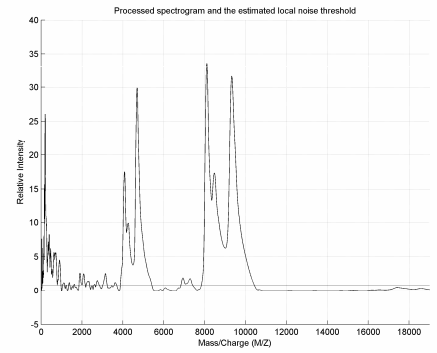

Figure 3. Global noise threshold estimation for each mass spectrum.

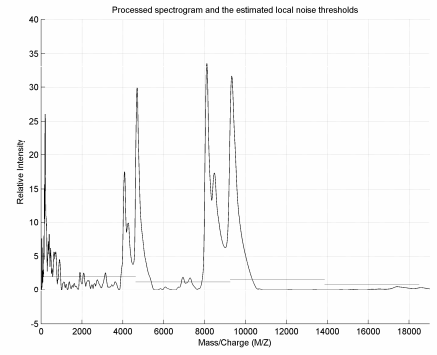

Figure 4. Local noise threshold estimation for each mass spectrum.

2.3 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 45-75 data points for each spectrum (see Fig 5). The varying number of peaks was due to chemical and electronic noise [13]. 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 ignoring the rest. At the end, an equal number of aligned peaks appeared in each mass spectrum.

Accordingly, the peak alignment method comprised the following steps:

1/ the local maxima (peaks) of each mass spectrum formed the spectrum's feature vector. The vector with the highest 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. $\boldsymbol{d_{min}}$ was selected to be 5 data points.

3/ $\boldsymbol{d_{min}}$ was used to form a $2*\boldsymbol{d_{min}}$ interval centred at the first peak of the reference vector. Within that interval, all feature vectors were scanned and if over than 10% 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 from the non-zero intensities of its corresponding initial spectra.

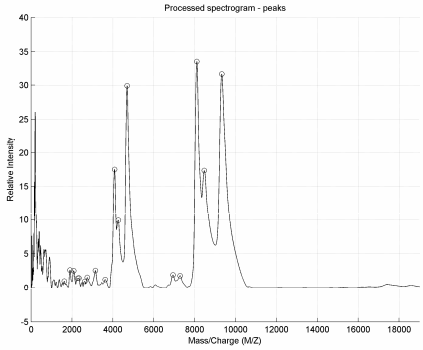

Figure 5. Local maxima (peaks) are specified with the circle symbol.

2.4 Classification

The classification tree structure comprised two levels (see Fig 6). Classification was performed by means of the Probabilistic Neural Network (PNN) classifier. The PNN determines each class probability density function (PDF) by linearly combining the kernel PDF estimation for each training sample separately for a given class. Its discriminant function is given by [14]:

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

where σ is the spread of the Gaussian activation function, 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.

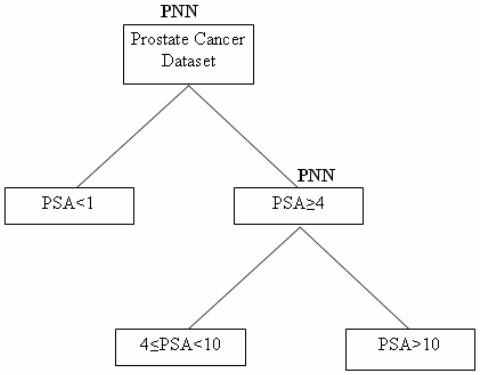

Figure 6. Prostate cancer classification tree structure..

According to the classification tree structure, the PNN was trained to discriminate controls with no evidence of disease (PSA<1) versus individuals with (PSA>4), and in the second node in distinguishing prostate cancer patients with PSA ranging between 4-10 versus patient with PSA value greater than 10. Feature sub-space selection was performed in order to retain only the feature subset with the highest discrimination information. For this purpose, the sequential backward feature selection algorithm was employed [15]. Systems' evaluation was performed by the standard procedure of the leave-one-out method (LOO) [15]. The classifier was designed

by all, leaving out one MS spectrum, which was then classified. The process was repeated, each time leaving-out a different 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, where the number of cases are few [4, 6, 8, 16-18].

3 RESULTS

Regarding the first node of the tree classification structure (see Fig.6), in distinguishing spectra with no evidence of disease (PSA<1) and spectra with prostate cancer (PSA≥4), PNN classifier scored 95.45% overall accuracy (see Table 1), utilizing the global noise estimation method. Employing the local noise estimation PNN classified correctly 97.7% of the samples (see Table 2). The feature (biomarkers) vectors using the two aforementioned noise estimation techniques are illustrated in table.3

PSA level	PSA<1	PSA≥4	Accuracy
PSA<1	60	3	95.24%
PSA≥4	3	66	95.65%
Overall accuracy			95.4%

Table 1 : PNN classification results, for 63 spectra with no evidence of disease (PSA<1) and 69 spectra with prostate cancer (PSA≥4) utilizing global noise estimation

PSA level	PSA<1	PSA≥4	Accuracy
PSA<1	61	2	96.83%
PSA≥4	1	68	98.55%
Overall accuracy			97.7%

Table 2 : PNN classification results, for 63 spectra with no evidence of disease (PSA<1) and 69 spectra with prostate cancer (PSA≥4) utilizing local noise estimation

Noise Estimation	Feature Vectors(m/z values)
Global	{1190.9, 1312, 1410.7, 1656.5, 1698, 1796, 2206.9, 3127.8, 3473.4, 3483.1, 5815.6, 6933.8, 8100.7, 9271.1, 10651, 11201.9, 11606, 12661.2}
Local	{1606.5, 1465.4, 1796, 2068.8, 2349.2, 3483.1, 4070.6, 6933.8, 7657.9, 8100.7, 8456.2, 9271.1, 10651, 11606}

Table 3 : Best biomarker vectors using global/local noise estimation, sequential backward selection, leave one out method and PNN in discriminating spectra with no evidence of disease (PSA<1) and spectra with prostate cancer (PSA≥4)

On the second node of the tree classification structure (see Fig.6) in discriminating prostate cancer spectra with PSA greater than 4 and less than 10 versus spectra with PSA greater than 10, PNN classifier scored 91.3% overall accuracy using the global noise estimation technique. Utilizing the local noise estimation PNN classified correctly 95.6% the unknown patterns. The feature (biomarkers) vectors using the local and global noise estimation techniques are depicted in table 6.

PSA level	4≤PSA<10	PSA>10	Accuracy
4≤PSA<10	22	4	84.62%
PSA>10	2	41	95.35%
Overall accuracy			91.3%

Table 4 : PNN classification results, for 26 spectra of prostate cancer with (4≤PSA<10) from 43 spectra of prostate cancer with (PSA>10) utilizing global noise estimation.

PSA level	4≤PSA<10	PSA>10	Accuracy
4≤PSA<10	25	1	96.15%
PSA>10	2	41	95.35%
Overall accuracy			95.6%

Table 5 : PNN classification results, for 26 spectra of prostate cancer with (4≤PSA<10) from 43 spectra of

prostate cancer with (PSA≥10) utilizing local noise estimation

Noise Estimation	Feature Vectors(m/z values)
Global	{1193.9, 1616.5, 2074, 2849, 4076.8 4400.6, 5257.4, 6253.6, 8105.6, 11580.8, 15264.3}
Local	{ 1137.2, 1381.4, 1517.1, 1699.1, 2159.8, 2200.5, 3149.4, 4260.4, 4675.6, 5824.5, 9322.7, 12180.3, 13414.8, 17428.5, 17727.3}

Table 6 : Best biomarker vectors using global/local noise estimation, sequential backward selection, leave one out method and PNN in distinguishing spectra with prostate cancer with (4≤PSA<10) from spectra of prostate cancer with (PSA>10)

4 DISCUSSION

The goal of this study was to propose biomarkers related to prostate cancer that discriminate between normal and cancerous cases and further between cases of elevated but different PSA values. Within this context, the contribution of two noise estimation techniques (global/local) was evaluated by means of a probabilistic neural network (PNN) classifier. The PNN classifier achieved 95.4% and 97.7% overall accuracies employing, local and global noise estimation respectively, in discriminating individuals with PSA<1 from those having PSA≥4, according to the first node of the classification tree structure (see Fig.6). For the second node of the classification tree (see Fig.6) utilizing global and local noise estimation methods, PNN scored 91.3% and 95.6% overall accuracies respectively in correctly distinguishing patients with 4≤ PSA<10 versus patients with PSA >10.

Previous studies [4, 5] employing the same dataset have also suggested biomarkers, which, however, differ among studies. This may be due to the various pre-processing schemes and classifiers employed by various studies. In the present study, seeking for proteins that might be related to prostate cancer, it was found, by searching in the ExPASy database [8], that the m/z value of 2068.8 (see Table3) was very close to the Nociceptin protein (2081.39 Daltons), which has been implicated in the stimulation of prostate cell growth [19], among other neuropeptides. The 2068.8 m/z value was detected employing the local noise estimation method. On the contrary, this value was missed utilizing the global noise estimation technique (see Table.3). Additionally, the 4675.6 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 [20], elevated levels of BCL-2 might aid to the progression of the prostate cancer. This biological marker (4675.6) was detected (see Table 6) by applying the local noise estimation method while it was missed by the global technique. Also, another m/z value (biomarker) derived employing the local noise estimation technique, was the 5824.5, 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 [21]. The present study revealed other m/z values as being significant in distinguishing prostate cancer dataset as illustrated in Tables 3 and 6. These values might constitute information rich biomarkers that have not been yet identified or related to prostate cancer.

5 ACKNOWLEDGEMENT

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] Chan, J.M., Stampfer, M.J., Giovannucci, E., Gann, P.H., Ma, J., Wilkinson, P., Hennekens, C.H., Pollak, M. (1998), "Plasma insulin-like growth factor-i and prostate cancer risk: A prospective study", Science, Vol. 279, pp. 563-566.

[2] McDavid, K., Lee, J., Fulton, J.P., Tonita, J., Thompson, T.D. (2004), "Prostate cancer incidence and mortality rates and trends in the united states and canada", Public Health Rep, Vol. 119, pp. 174-186.

[3] 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, G.L., Jr. (2002), "Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men", Cancer Res, Vol. 62, pp. 3609-3614.

[4] Jong, K., Marchiori, E., Sebag, M., van der Vaart, A. (2004), "Feature selection in proteomic pattern data with support vector machines", *Proceedings of* 41.
[5] Lilien, R.H., Farid, H., Donald, B.R. (2003), "Probabilistic disease classification of expression-dependent proteomic data from mass spectrometry of human serum", J Comput Biol, Vol. 10, pp. 925-946.
[6] 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., Feng, Z. (2003), "Data reduction using a discrete wavelet transform in discriminant analysis of very high dimensionality data", Biometrics, Vol. 59, pp. 143-151.
[7] 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. (2002), "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. 1835-1843.
[8] 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. (2003), "A data-analytic strategy for protein biomarker discovery: Profiling of high-dimensional proteomic data for cancer detection", Biostatistics, Vol. 4, pp. 449-463.
[9] Institute, N.C. http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp, 24/11/2006.
[10] Hilario, M., Kalousis, A., Pellegrini, C., Muller, M. (2006), "Processing and classification of protein mass spectra", Mass Spectrom Rev, Vol. 25, pp. 409-449.
[11] Cleveland, W.S. (1979), "Robust locally weighted regression and smoothing scatterplots", J Amer Statist Assoc, Vol. 74, pp. 829–836.
[12] Wang, X., Zhu, W., Pradhan, K., Ji, C., Ma, Y., Semmes, O.J., Glimm, J., Mitchell, J. (2006), "Feature extraction in the analysis of proteomic mass spectra", Proteomics, Vol. 6, pp. 2095-2100.
[13] Baggerly, K.A., Morris, J.S., Coombes, K.R. (2004), "Reproducibility of seldi-tof protein patterns in serum: Comparing datasets from different experiments", Bioinformatics, Vol. 20, pp. 777-785.
[14] Specht, D.F. (1990), "Probabilistic neural networks", Neural Networks, Vol. 3, pp. 109-118.
[15] Theodorides, S., Koutroumbas, K. (2003), *Pattern recognition.*
[16] Lancashire, L., Schmid, O., Shah, H., Ball, G. (2005), "Classification of bacterial species from proteomic data using combinatorial approaches incorporating artificial neural networks, cluster analysis and principal components analysis", Bioinformatics, Vol. 21, pp. 2191-2199.
[17] Ressom, 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. (2005), "Analysis of mass spectral serum profiles for biomarker selection", Bioinformatics, Vol. 21, pp. 4039-4045.
[18] Sorace, J.M., Zhan, M. (2003), "A data review and re-assessment of ovarian cancer serum proteomic profiling", BMC Bioinformatics, Vol. 4, pp. 24.
[19] Swanson, T.A., Kim, S.I., Myers, M., Pabon, A., Philibert, K.D., Wang, M., Glucksman, M.J. (2004), "The role of neuropeptide processing enzymes in endocrine (prostate) cancer: Ec 3.4.24.15 (ep24.15)", Protein Pept Lett, Vol. 11, pp. 471-478.
[20] Hering, F.L., Lipay, M.V., Lipay, M.A., Rodrigues, P.R., Nesralah, L.J., Srougi, M. (2001), "Comparison of positivity frequency of bcl-2 expression in prostate adenocarcinoma with low and high gleason score", Sao Paulo Med J, Vol. 119, pp. 138-141.
[21] Pan, C.X., Kinch, M.S., Kiener, P.A., Langermann, S., Serrero, G., Sun, L., Corvera, J., Sweeney, C.J., Li, L., Zhang, S., Baldridge, L.A., Jones, T.D., Koch, M.O., Ulbright, T.M., Eble, J.N., Cheng, L. (2004), "Pc cell-derived growth factor expression in prostatic intraepithelial neoplasia and prostatic adenocarcinoma", Clin Cancer Res, Vol. 10, pp. 1333-1337.