Computational Intelligence Solutions for Biomarker Discovery in Mass Spectrometry Data.

Panagiotis Bougioukos[1], Dionisis Cavouras[2], Antonis Daskalakis[1], George Nikiforidis[1], Sofia Kossida[3], Antonia Vlahou[3] and Anastasios Bezerianos[1]

[1] Department of Medical Physics, School of Medicine, University of Patras, 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

[3] Division of Biotechnology, Foundation for Biomedical Research, Academy of Athens, Athens, Greece.

Introduction

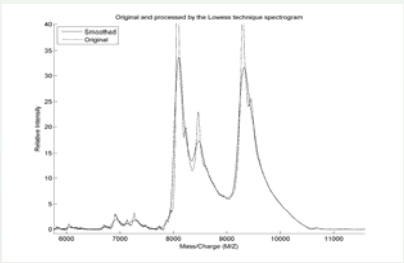
Early detection of cancer is a critical issue for improving patient survival rates. Prostate cancer is a very common cancer disease. The most widely used method for prostate cancer detection is measuring the concentration of the prostate specific antigen (PSA). PSA is the best marker used in clinical practice. The method has the desirable property of yielding high sensitivity but also the drawback that specificity is relatively low[1]. Recently studies indicates that an alternative approach to prostate cancer detection is provided through the analysis of data (Spectra) obtained from Surface Enhanced Laser Desorption Ionization (SELDI) mass spectrometry (MS) or Matrix Assisted Laser Desorption Ionization (MALDI) mass spectrometry (MS). The end-result of the MS data processing and analysis pipeline[2] is a set of differentially expressed or altered proteins (biomarkers) that potentially influences prostate cancer diagnosis.

In the present study, an alternative approach to identify possible potential biomarkers is proposed. Accordingly, information rich features (biomarkers) were found that improved the performance of a Probabilistic Neural Network (PNN) classifier in differentiating among datasets with PSA level greater than 10, of prostate cancer and with no evidence of disease.

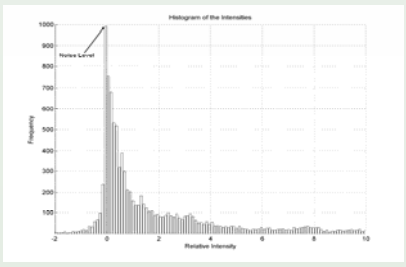
Materials and Methods

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, 63 spectra with no evidence of disease (PSA<1) and 43 spectra with prostate cancer (PSA>10).

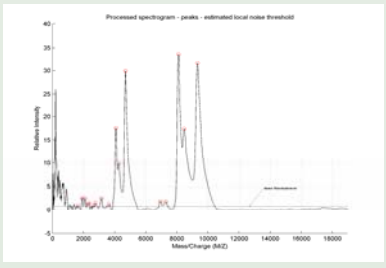
Smoothing: Signal noise contamination was reduced by the lowess smoothing technique (see figure below)



Thresholding: 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.



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.



The varying number of peaks due to chemical and electronic noise was alleviated, developing a peak alignment process, 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.

Classification: 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 :

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

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 kth pattern vector of class i.

Results

Classification results for 63 spectra with no evidence of disease (*PSA<1*) and 43 spectra with prostate cancer (*PSA>10*).

No Spectra / PSA level	PSA<1	PSA>10	Accuracy
63 / PSA<1	62	1	98,4%
43 / PSA>10	1	42	97,6%

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.

Accuracies obtained by the proposed method demonstrate that SELDI protein chip mass spectrometry combined with a PNN classification algorithm facilitate the determination of informative biomarkers for prostate cancer providing an innovative clinical diagnostic platform.

References

[1] Banez, L. L.,et.all:"Proteomics in prostate cancer". Curr Opin Urol. 15 (2005) 151-6

[2] Petricoin, E. F.et.all :"Serum proteomic patterns for detection of prostate cancer." J Natl Cancer Inst. 94 (2002) 1576-8

[3] Specht, D.F. (1990) "Probabilistic neural networks," Neural Networks, 3,110-118.

Acknowledgements

This work is co-financed by 75% from European Union-European Social Fund and by 25% from National Fund (013/PENED03).