International Journal of Medical, Medicine and Health Sciences
ISSN: 2517-9969
Vol:4, No:12, 2010

Diagnosis of Ovarian Cancer with Proteomic Patterns in Serum using Independent Component Analysis and Neural Networks

Simone C. F. Neves, Lúcio F. A. Campos, Ewaldo Santana, Ginalber L. O. Serra, Allan K. Barros

***Abstract*—**We propose a method for discrimination and classification of ovarian with benign, malignant and normal tissue using independent component analysis and neural networks. The method was tested for a proteomic patters set from A database, and radial basis functions neural networks. The best performance was obtained with probabilistic neural networks, resulting I 99% success rate, with 98% of specificity e 100% of sensitivity.

***Keywords*—**Cancer ovarian, Proteomic patterns in serum, independent component analysis and neural networks.

I. INTRODUCTION

OVARIAN cancer is the most lethal gynecologic malignancy. Poor survival rates are mainly attributable to late diagnosis. For instance, the widely used biomarker of cancer antigen 125 (CA125) for ovarian cancer can only detect 50-60% of patients with stage I ovarian cancer while the positive predictive value less than 10%. Clearly, there is urgent need to unravel novel biomarkers for early detection of ovarian cancer. New proteomic technologies have brought the hope of discovering novel early cancer-specific biomarkers in complex biological samples. Novel mass spectrometry (MS) based technologies in particular, such as surface-enhanced laser desorption/ionisation time of flight (SELDI-TOF-MS), have shown promising results in recent years. SELDI-TOF MS combined with bioinformatics approach has successfully found some new biomarkers and achieved high sensitivity and specificity for the diagnosis of ovarian cancer [1]–[3], breast cancer [4], prostate cancer [5], [6], colorectal cancer [7], lung cancer [8] and so on. Cancer detection based on the application of feature extration techniques to proteomic data has received a lot of attention in recent years [9]–[11]. The mass spectrum data present a curve with peaks and valleys, where the x-coordinate is the ratio of molecular weight to the net electrical charge for a specific organic molecule, with Dalton (Da) as unit, and biomarker identification. Although proteomic mass spectra has shown the promising potential of finding disease-related protein patterns, key challenges remain in the processing of them especially for the curse of dimensionality. In the present study, an alternative approach to feature extraction from ICA data of ovarian cancer is proposed.

Simone Neves is with Federal University of Maranhão Brazil. e-mail:neves.simone@gmail.com

Petricoin et al. [1] combined a genetic algorithm with selforganizing cluster analysis for identifying ovarian cancer. They reported an discriminatory pattern for ovarian cancer, which was defined by the amplitudes at five key m/z values. A sensitivity of 100%, with 95% confidence interval of 93-100%, and a specificity of 95%, with 95% confidence interval of 87-99% were reported. Adam et al. [9] applied decision-tree learning to mass spectra of prostate cancer patients. They used Ciphergen SELDI (r) software for peak detection, and decision trees for classification using the intensity levels of the nine highest discriminatory peaks as features. This technique gave 96% accuracy, 83% sensitivity and 97% specificity. Ball et al. [11] applied a three-layer perceptron artificial neural network (ANN) (Neuroshell 2) with a back propagation algorithm to analyze mass spectra for predicting astroglial tumor grade (1or 2). Relative intensity patterns were significantly reduced in high-grade astrocytoma. The accuracy achieved was between 83 and 100% for predicting tumor grade, however, the sample size for this study was only 12.

This paper introduces the method of cancer markers pattern analysis, which based on feature extraction using independent component analysis and classification with neural networks, establishing a new pattern for diagnosis of ovarian cancer.

II. METHODS

Let $s(t)$ be an proteomic signal and let us take $s(t)$ in m windows of fixed length τ, where each windows is a sample from a patient.

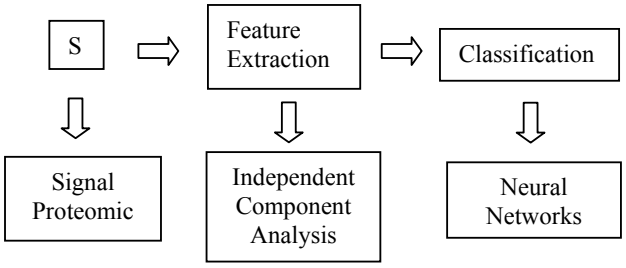

Fig. 1. Block diagram of the proposed method

The block diagram of the proposed method is shown in Figure 1. It consists of the extraction of features using ICA and classification of the through neural networks.

2.1 Independent Component Analysis

Let us assume that we have *n* random variables *s*1,.., *s*n (random streams of an patient signal), modeled as linear combinations of *n* random variables *a*1,…, *a*n, such that

$$s = \varphi_{i1} a_1 + \cdots + \varphi_{in} a_n \text{ for all i = 1,…,n} \quad (1)$$

where φ in are real coefficients. Define **s**, ***a*** and **Φ** as.

$$s = [s_1 \quad s_2 \quad s_n]^T \quad (2)$$

$$\Phi = \begin{bmatrix} \varphi_{11} & \cdots & \varphi_{1n} \\ \vdots & & \vdots \\ \varphi_{n1} & \cdots & \varphi_{nn} \end{bmatrix} \quad (3)$$

$$a = [a_1 \quad a_2 \quad a_n]^T \quad (4)$$

Using (2), (3) and (4) to rewrite (1), we obtain

$$s = \Phi a \quad (5)$$

The objective of ICA is to estimate the matrix **Φ** so that the a_i are statically mutually independent. Let **B** be the inverse of **Φ**. Then we can state that

$$a = \mathrm{B}s \quad (6)$$

2.1.1 The FastICA Algorithm

We use the FastICA algorithm to find the matrix **B**. Hyvärinen *et al* [12] summarizes the algorithm in the following 7 steps:

1. Center the data and make its mean zero.
2. Whiten the data to give **z**.
3. Choose *m*, the number of independent components to estimate.
4. Choose initial values for the **b**i, *i* = 1,…, *m*, each of unit norm. Orthogonalize the matrix **B** as in step 6 below.
5. For every *i* = 1,…, *m*, let **b**i ← E{**z**g(**b**i′**z**)}-E{g′(**b**i′**z**)}**w**, where g is defined.
6. Do a symmetric orthogonalization of the matrix **B** = (**b**1,…,**b**m)′ by **B**←(**BB**′)$^{-1/2}$**B** (7)
7. If not converged, go back to step 5.

After the estimation of **B**, we can easily obtain **Φ**. We are interested in the columns of **Φ,** which are called *basis functions* of **s**.

2.2 Neural Network

In this work, we use a Multilayer Perceptron Neural Network (MLP), Probabilistic Neural Network (PNN) and Radial Basis Functions Neural Network (RBFNN) to classify malignant, benign and normal tissues.

2.2.1 Multilayer Perceptron Neural Networks

The Multilayer Perceptron (MLP), a feed-forward back-propagation network, is the most frequently used neural network technique in pattern recognition [13], [14]. Speaking, MLPs are supervised learning classifiers that consist of an input layer, an output layer, and one or more hidden layers that extract useful information during learning and assign modifiable weighting coefficients to components of the input layers. In the first (forward) pass, weights assigned to the input units and the nodes in the hidden layers and between the nodes in the hidden layer and the output, determine the output. The output is compared with the target output. An error signal is then back propagated and the connection weights are adjusted correspondingly. During training, MLPs construct a multidimensional space, defined by the activation of the hidden nodes, so that the three classes (malignant, benign and normal tissue) are as separable as possible. The separating surface adapts to the data.

2.2.2 Probabilistic Neural Network

The probabilistic neural network (PNN) is a direct continuation of the work on Bayes classifiers. The PNN learns to approximate the *pdf* of the training examples [14]. More precisely, the PNN is interpreted as a function which approximates the probability density of the underlying example The PNN consists of nodes allocated in three layers after the inputs:

- *pattern layer*: there is one pattern node for each training example. Each pattern node forms a product of the weight vector and the given example for classification, where the weights entering a node are from a particular example. After that, the product is passed through the activation function:

$$\exp\left[\left(x^T w_{Ki} - 1\right)/\sigma^2\right] \quad (8)$$

Where
x: Data input
W_k: Weight
σ: Smothing adjust

- *summation layer*: each summation node receives the outputs from pattern nodesassociated with a given class:

$$\sum_{i=1}^{Nk} \exp\left[\left(x^T w_{Ki} - 1\right)/\sigma^2\right] \quad (9)$$

-*output layer*: the output nodes are binary neurons that produce the classification decision

$$\sum_{i=1}^{Nk} \exp\left[\left(x^T w_{Ki} - 1\right)/\sigma^2\right] > \sum_{i=1}^{Nj} \exp\left[\left(x^T w_{Ki} - 1\right)/\sigma^2\right] \quad (10)$$

2.2.3 Radial Basis Functions Neural Network

Successful implementation of the Radial Basis Functions Neural Network (RBFNN) can be achieved using efficient supervised or unsupervised learning algorithms for an accurate estimation of the hidden layer [15]-[16].

In our implementation, the k-means unsupervised algorithm was used to estimate the hidden layer weights from a set of training data containing the features from malignant, benign and normal tissue. After the initial training and the estimation of the hidden layer weights, the weights in the output layer are computed using Wiener filter, for example, by minimizing the mean square error (MSE) between the actual and the desired output over the set of samples.

The RBFNN have a faster learning rate and have been proved to provide excellent discrimination in many applications.

2.3 Selection of Most Significant Features

Our main objective is to identify the effectiveness of a feature or a combination of features when applied to a neural network. Thus, the choice of features to be extracted is important.

Forward selection is a method to find the "best" combination of features (variables) by starting with a single feature, and increasing the number of used features, step by step [17]. In this approach, one adds features to the model one at a time. At each step, each feature that is not already in the model is tested for inclusion in the model. The most significant of these feature is added to the model, so long as P-value is below some pre-selected level.

2.4 Evaluation of the Classification Method

Sensitivity and specificity are the most widely used statistics to describe a diagnostic test. Sensitivity is the proportion of true positives that are correctly identified by the test and is defined by ***S = TP/(TP+FN).*** Specificity is the proportion of true negatives that are correctly identified by the test and is defined by ***TN/(TN+FP).*** Where **FN** is false-negative, **FP** is false-positive, **TN** is true negative and **TP** is true positive diagnosis.

III. EXPERIMENTAL RESULTS AND DISCUSSIONS

Here are describe the results obtained using the method proposed in the previous section.

3.1 Proteomic Patterns Database

The serum SELDI MS data sets were used in this research to identify serum proteomic patterns that distinguish the serum of ovarian cancer cases from non-cancer controls. The data sets were downloaded from a public website: http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp. As explained on the website, dataset (Ovarian, 04-03-02) consisted of 100 cancerous samples, 100 non-cancerous samples, and 16 benign samples. Only the cancerous and non-cancerous samples are included in this paper. Each sample consisted of 15,154 intensities corresponding to 15,154 m/z values with intensities of features.

3.2 ICA Applications

Each sample represents one row of the mixture matrix. The matrix **S** is represented by the samples into the dimension of **P**, that is, 1x15,154. Thus, each row of the matrix **B** correspond to a functions basis, and each column correspond to an attributed weight to a intensity value, i.e., an x input parameter to the neural network [18]. Using the FastICA algorithm and the matrix ***S***, we obtain the basis function matrix **B**, which contains the features of each sample.

3.1 Neural Networks

Using the *forward-selection* algorithm, basis functions were selected as being the most significant features. The chosen features (a_i) are the input to the Neural Network. For each Neural Networks (MLP, PNN, RBFNN) the algorithm selected the most significant features.

We carried out tests with different Neural Network architetures to find the bests MLP, PNN and RBF Neural Networks.

In order to carry out the tests, we divided a sample in 128 samples patients: 64 for training and 64 for tests.

3.1 Results

Table 1 shows the Neural Networks of the application of the ICA technique with each Neural Network for discrimination patterns.

TABLE I NEURAL NETWORKS ARCHITECTURE AND CLASSIFICATION OF MALIGNANT, BENIGN AND NORMAL

	TP	TN	FP	FN	Specificity	Sensitivity	Accuracy
PNN	64	62	2	0	97%	100%	98%
RBF	64	63	1	0	98%	100%	99%
MLP	63	63	1	1	98%	98%	98%

Based on the Table 1, the best results was obtained with Probabilistic Neural Networks. The RBF obtained a success rate of 99 % on discriminating malignant, benign and normal tissues. The found specificity was 98% and the sensitivity, 100%. The RBF obtained 126 true positives diagnosis, 0 true negatives, 1 false positives.

IV. CONCLUSION

The presented results demonstrate that Independent Component Analysis and Neural Networks is a useful tool to discriminate malignant, benign and normal tissues. Furthermore, the Probabilistic Neural Network obtained the best performance, classifying those tissues, with a success rate of 99%, specificity of 98 % and sensitivity of 100%. It can decrease the number of unneeded biopsies and late cancer diagnosis. Based on these results, we have observed that such features provide significant support to a more detailed clinical investigation, and the results were very encouraging when tissues were classified with ICA and Neural Networks.

ACKNOWLEDGMENT

To all my co-workers in the Laboratory for Biological Information Processing, specially Lúcio Flávio and Cristiane Silva.

REFERENCES

[1] E. F.Petricoin, "Use of proteomic patterns serum to identify ovarian cancer", *The Lancet*, vol 359, pp 572-577, 2002.

[2] K.R.Kozak, "Characterization of serum biomarkers for detection od early stage ovarian cancer,"*Proteomics*, vol 17,no. 5, pp.4589-4596, September 2005.

[3] J.K.Yu, "An integrated approach utilizing proteomics and bioinformatics to detect ovarian cancer," *Jornal of Zhejiang University SCIENCE,* vol. 6, no.4, pp. 227-231, July 2005.

[4] G. Ricolleau, " Suface – enhaced laser desorption/ionization time of flight mass spectrometry protein profiling identifies ubiquitin and ferritin light chain as prognostic biomarkers in node-negative breast cancer tumors." *Proteomics*, vol.6, no.6, pp.1963-1975, November 2006.

[5] D. Donald. "Bagged super wavelets reduction for boosted prostate cancer classification of seldi-tof mass spectral serum profiles, "*Chemometrics and intelligent Laboratory Systems*, vol 82, no. 1, pp. 2-7, January 2006.

[6] W.D. Tong, "Using decision forest to classify prostate cancer sample son the basis of seldi-tof ms data: assessing chance correlation and prediction confidence," *Enviromental Health perspective*, vol 112, no.16, November 2004.
[7] J. K. Yu, "An integrated approach to the detection of colorectal cancer utilizing proteomics and bioinformatics," *World J Gastroenterol*, vol.21, no.10, pp. 3127-3131, October 2004.
[8] S. Y. Yang, "Application of serum seldi proteomic patterns in diagnosis of lung cancer, " *BMC cancer*, vol. 83, no. September 2005.
[9] B. L. Adam, "Serum protein finger printing coupled with a pattern matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and heatthy men ", *Cancer Res*, vol.62,no.3, pp.609-614,May 2002.
[10] Y. Qu., "Boosted decision tree analysis of surface – enhaced laser desorption/ionization mass spectral serum profiles discriminated prostate cancer from nonprostate patients," *Clin Chem*, vol.48, pp.1835-1843, 2002.
[11] G. Ball, "An integrated approach utilizing artificial neural networks and seldi mass spectrometry for the classification of human tumors and rapid dentification of potencial biomarkers," Some Fine Journal, vol.18, no. 3, pp. 395 – 4004, May 2002.
[12] Hyvärinen, A., J. Karhunen, and E. Oja, Independent Component Analysis, *John Wiley & Sons*, New York, 2001.
[13] R. O. Duda, "Pattern Classification and Scene Analysis," *Wiley Intercience Publication*, New York, 1973.
[14] C. M. Bishop, "Neural Networks for Pattern Recognition," *Oxford University Press*, New York, 1999.
[15] Christoyianni I., "Fast detection of masses in computer-aided mammography,". *IEEE Signal Process Mag 2000*, no. 7, pp 54-64.
[16] Christoyianni I., "Neural classification of abnormal tissue in digital mammography using statistical features of the texture," *IEEE Int Conf Electron*, Circuits Systems, no.1, pp. 117-120, 1999.
[17] K. Fukunaga, "Introduction to Statistical Pattern Recognition," *Academic Press*, London, 1990.
[18] Christoyianni I., "Computer aided diagnosis of breast cancer in digitized mammograms,",*Comp. Med. Imag. & Graf*., no. 26, pp. 309-319, 2002.