IJISET - International Journal of Innovative Science, Engineering & Technology, Vol. 3 Issue 1, January 2016.
www.ijiset.com
ISSN 2348 – 7968

Identifying Significant Features of Mass Spectra Samples for Ovarian Cancer Detection

Anjali Sharma[1] and Satnam Singh[2]

[1] M. Tech Scholar, Department of ECE, Sri Sai College of Engineering & Technology, Badhani, Punjab, India
[2] Asstt. Professor, Department of ECE, Sri Sai College of Engineering & Technology, Badhani, Punjab, India
[1] engineeranjalisharma@gmail.com

Abstract

Cancer detection research is an interesting research area in the field of medical science. Early detection and classification is necessary for cancer diagnosis and treatment. Formerly cancer classification depends upon the morphological and clinical however the modern arrival of the micro array technology has permitted the concurrent observation of thousands of genes, which provoked the progress in cancer classification using gene expression data. Therefore, the diagnosis of early stage ovarian tumour would significantly decrease the morbidity and mortality rate from this disease. The aim of this paper is to review selectively recent progress in the pathology of ovarian tumours. This paper classifies mass spectrometry data and shows some statistical tools that can be used to look for potential disease markers and proteomic pattern diagnostics.

Keywords: *Ovarian cancer, C*A-125, SELDI, serum, DNA

1. Introduction

In 2012, ovarian cancer occurred in 239,000 women and resulted in 152,000 deaths worldwide. This makes it, among women, the seventh-most common cancer and the eighth-most common cause of death from cancer. Death from ovarian cancer is more common in North America and Europe than in Africa and Asia.

Ovarian tumor is the most common and fifth most common cause of death in women. The American Cancer Society estimated that there would be 25,400 new cases of ovarian tumor and 20,000 deaths. Since the 1960s, almost 80% of women with ovarian tumor are diagnosed when the disease has spread to the upper abdomen (stage III) or beyond (stage IV). Unfortunately, the 5-year survival rate for those women is approximately 15%, whereas the 5-year survival when detected at early stage (I) approaches 90%. Therefore, the diagnosis of early stage ovarian tumor would significantly decrease the morbidity and mortality rate from this disease. In order to find the cure it is necessary to quickly diagnose the disease accurately and treat it based on the kind of symptoms appeared. Ovarian tumor has several classifications, which may help to determine the best treatment. Ovarian cancer begins in the ovaries. Ovaries are reproductive glands found only in females (women). The ovaries produce eggs (ova) for reproduction. The eggs travel through the fallopian tubes into the uterus where the fertilized egg implants and develops into a fetus. The ovaries are also the main source of the female hormones estrogen and progesterone [1]. One ovary is on each side of the uterus in the pelvis. The ovaries are made up of 3 main kinds of cells. Each type of cell can develop into a different type of tumor:

Epithelial tumors start from the cells that cover the outer surface of the ovary. Most ovarian tumors are epithelial cell tumors.

Germ cell tumors start from the cells that produce the eggs (ova).

Stromal tumors start from structural tissue cells that hold the ovary together and produce the female hormones estrogen and progesterone.

Serum proteomic profiling, by using surfaced-enhanced laser desorption mass spectrometry is one of the most promising new techniques for cancer diagnostics. Exceptional sensitivities and specificities have been reported for some cancer types such as prostate, ovarian, breast, and bladder cancers [2]. These sensitivities/specificities are far superior to those obtained by using classical cancer biomarkers.

2. Mass Spectrometry (MS)

Mass spectrometry has been widely used as a diagnostic tool in clinical laboratories. It has been used with success for the identification & quantification of small molecules (molecular mass < 1,000 Da). These molecules could be highly informative in:

(a) new-born screening programs (39)
(b) Toxicological and forensic applications (40).
(c) For delineating various types of inborn errors of metabolism (41)
(d) For detecting doping of athletes (42), etc.

ISSN 2348 – 7968

Over the last 15 years, we have seen a revival of this technology for studying larger molecules such as nucleic acids and proteins. These new applications became possible mainly due to the development of novel methodologies to effectively volatize and ionize proteins and nucleic acids, by using various chemicals (matrices) and lasers (e.g. matrix-assisted laser desorption/ionization, MALDI) or electro-spray ionization (ESI). The ability to measure with high accuracy mass-to-charge ratio, providing spectra of very high resolution, and the development of tandem mass spectrometry (MS/MS) to obtain de novo protein sequence information has further enhanced the applications of this technology in proteomics [3]. Coupling of mass spectrometers to liquid chromatro-graphy (LC/MS) further expanded the discriminatory power of the method.

Mass spectrometry is now one of the most powerful proteomic tools [4]. Even more spectacular advances in mass spectrometry should be expected, with further improvements in resolution and detectability. With this in mind, it is not surprising that many scientists have decided to use mass spectrometry either as a diagnostic tool or as a cancer or other disease biomarker discovery platform [5]. These methods and fields of investigation, used appropriately, may indeed succeed in discovering new diagnostic modalities for cancer and other diseases, as well as contribute to the better understanding of the pathogenesis of such diseases. The Human Proteome Organization (HUPO, www.hupo.org) is focusing on the identification of large numbers of proteins in complex mixtures, including serum and other biological fluids. It is expected that these efforts will finally lead to the identification of new potential biomarkers for cancer and other diseases. HUPO also intends to standardize the methodology so that the results obtained with these techniques are robust and reproducible among laboratories. Most of the discussion below will focus on one proteomic platform used extensively in diagnostics, known as surface enhanced laser desorption/ionization-time-of-flight (SELDI-TOF) mass spectrometry. This technique is based on the pre-treatment of a biological fluid or tissue extract with various proteomic chips, performing protein extractions based on hydrophobic, ion-exchange, metal binding, or other interactions.

3. MS as a Cancer Biomarker

This approach represents a paradigm shift in cancer diagnostics, based on complex mass spectrometric differences between proteomic patterns in serum between patients with or without cancer identified by bioinformatics. Their premise is that no matter what the nature of these molecules are, their potential to discriminate between these two conditions should be further exploited. The central hypothesis of this approach is as follows: protein or protein fragments produced by cancer cells or their microenvironment may eventually enter the general circulation. Then, the concentration (abundance) of these proteins/fragments could be analysed by mass spectrometry and used for diagnostic purposes, in combination with a mathematical algorithm. The vast majority of the currently available data have been produced by using the SELDI-TOF technology, marketed by Ciphergen Biosystems (Fremont, CA). Ciphergen claims that over 200 papers have already been published with this technology.

The types of cancers that have been examined include ovarian, prostate, breast, bladder, renal, and others, and the biological fluids analysed include serum, urine, cerebrospinal fluid, nipple aspirate fluid, etc. The apparent successes with this technology have been recently reviewed by many investigators. In general, it has been suggested that this technology can achieve much higher diagnostic sensitivity and specificity (approaching 100%) in comparison to the classical cancer biomarkers [6]. The technology's potential has been expanded to other diseases such as Alzheimer's disease, Creutzfeldt-Jakob disease, renal allograft rejection, etc.

The analytical procedure with this technology involves a few common steps. The biological fluid of interest is first interacted with a protein chip that incorporates some kind of an affinity separation between "non-informative" and "informative" proteins. After washing, the immobilized (and fortunately mostly informative) proteins can be studied by using SELDI-TOF mass spectrometry. Two types of data have been reported in the literature: 1) discriminating peaks of unknown identity that are different in amplitude (increased or decreased) between normal individuals and patients with cancer; and 2) data in which at least some of these peaks have been positively identified (see below). Computer algorithms have been used to analyze these multidimensional data to demonstrate that a pattern consisting of several peaks (from tens to thousands) is sufficiently different between the two groups of subjects. This technology is now seen as the most promising way of diagnosing early cancer [7]. Clinical trials are now underway and will reveal, in a blinded fashion, if these data can be reproduced and if they are robust enough for clinical use. The use of SELDI-TOF technology as a cancer biomarker discovery tool (as opposed to a cancer diagnostic tool) is straightforward. The discriminatory peaks, if positively identified, may represent molecules that could be measured with simpler and cheaper techniques for the purpose of diagnosing cancer. For example, some investigators postulate that

ISSN 2348 – 7968

such molecules may be routinely quantified by using enzyme-linked immune-sorbent assay (ELISA) technologies [8]. In practice, very few, if any, of the SELDI-TOF identified novel candidate biomarkers have been validated by using alternative technologies.

4. Mass Spectrometry Curve

The development of tools for the early cancer diagnosis is a major open problem, and clinicians have investigated a variety of diagnosis techniques. Recently, they have discovered that cancer may affect the blood mass spectrum, and studied diagnosis methods based on the analysis of mass-spectrum data, which provide information about proteins and their fragment [9]. The blood mass spectrum is a curve (Figure 1), where the x-axis shows the ratio of the weight of a specific molecule to its electric charge, and the y-axis is the signal intensity for the same molecule.

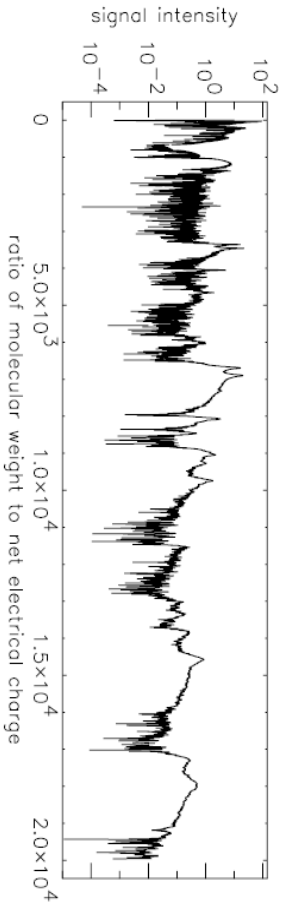

Figure 1: Mass-spectrum curve.

The mass-spectrum analysis is a fast inexpensive procedure based on a sample of a patient's blood, and it may potentially allow cancer screening with little discomfort to a patient. Medical researchers have developed techniques for the detection of early cancer based on protein markers, which are certain molecules in body tissues and fluids, but these techniques are often inaccurate. Recently, researchers have developed a new cancer-detection method based on the application of data mining to the mass spectra of patients' tissue cells, blood, serum, and other body fluids

Medical researchers have developed several techniques for analyzing the mass-spectrum data, which allow the diagnosis of various cancers, including ovarian, breast, prostate, bladder, pancreatic, kidney, liver, and colon cancers. The effectiveness of these techniques varies across cancer types, methods for generating mass spectra, and algorithms for analysing the resulting data. Clinicians use three standard measures of the effectiveness of diagnosis techniques: sensitivity, specificity, and accuracy. The sensitivity is the probability of the correct diagnosis for a patient with cancer, the specificity is the chances of the correct diagnosis for a healthy person, and the accuracy is the chances of the correct diagnosis for the overall population of healthy and sick people [10].

5. Identifying Significant Feature

Serum proteomic pattern diagnostics is used to distinguish samples from patients with and without cancer. Profiles patterns are generated using surface enhanced laser desorption and ionization called SELDI protein mass spectrometry. We pre-process the raw dataset/spectrogram in the files and returns the mass/charge (MZ) and ion intensities (Y) vectors. Each column in 'Y' represents measurements taken from a patient. There are 216 columns in 'Y' representing 216 patients, out of which 121 are ovarian cancer patients and 95 are normal patients. Each row in 'Y' represents the ion intensity level at a specific mass-charge value indicated in 'MZ'. There are '15000' mass-charge values in 'MZ' and each row in 'Y' represents the ion-intensity levels of the patients at that particular mass-charge value. The variable 'grp' holds the index information as to which of these samples represent cancer patients and which ones represent normal patients. The goal is to select a reduced set of measurements or "features" that can be used to distinguish between cancer and control patients. These features will be ion intensity levels at specific mass/charge values. We plot some data sets into a figure 2 window to visually compare profiles from the two groups; five spectrograms from cancer patients (blue) and five from control patients (green) are displayed zooming in on the region from 8500 to 8700

ISSN 2348 – 7968

M/Z shows some peaks that might be useful for classifying the data.

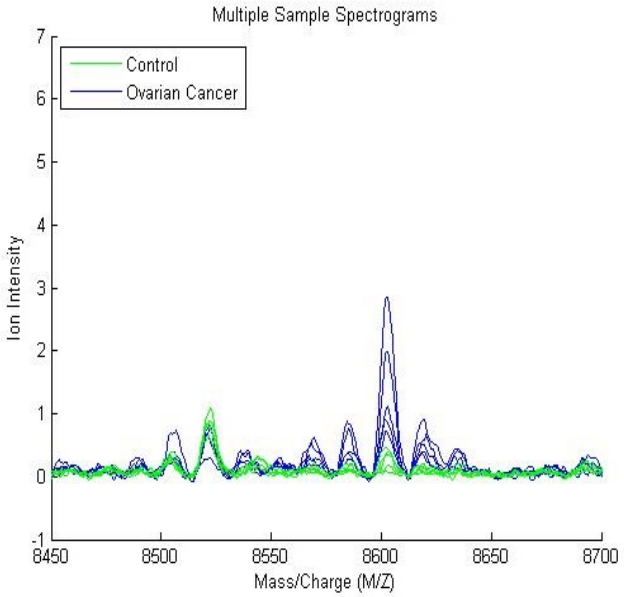

Figure 2: Five spectrograms of Cancer & Normal patients

Another way to visualize the whole data set is to look at the group average signal for the control and cancer samples. You can plot the group average and the envelopes of each group. Figure 3 shows the averages of two groups.

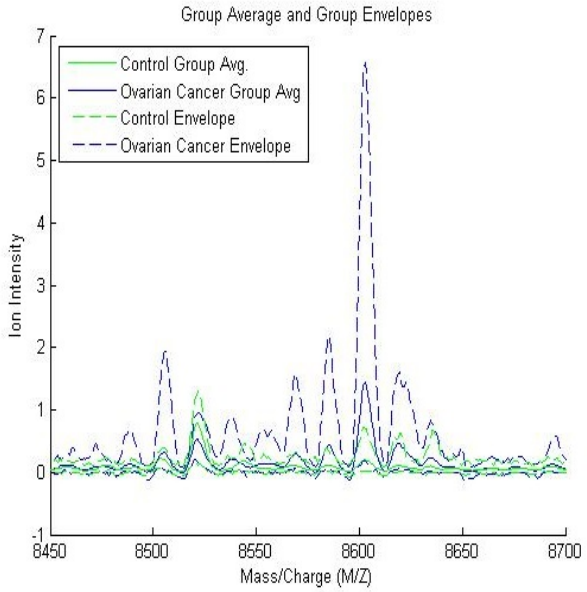

Figure 3: Group Average of control and cancer samples.

Observe that apparently there is no single feature that can discriminate both groups perfectly. A simple approach for finding significant features is to assume that each M/Z value is independent and compute a two-way t-test. Rank-features function returns an index to the most significant M/Z values, for instance 100 indices ranked by the absolute value of the test statistic. This feature selection method is also known as a filtering method, where the learning algorithm is not involved on how the features are selected. The first output given below of rank-features can be used to extract the M/Z values of the significant features. The second output of rank-features is a vector with the absolute value of the test statistic. We plot it over the spectra using plot function.

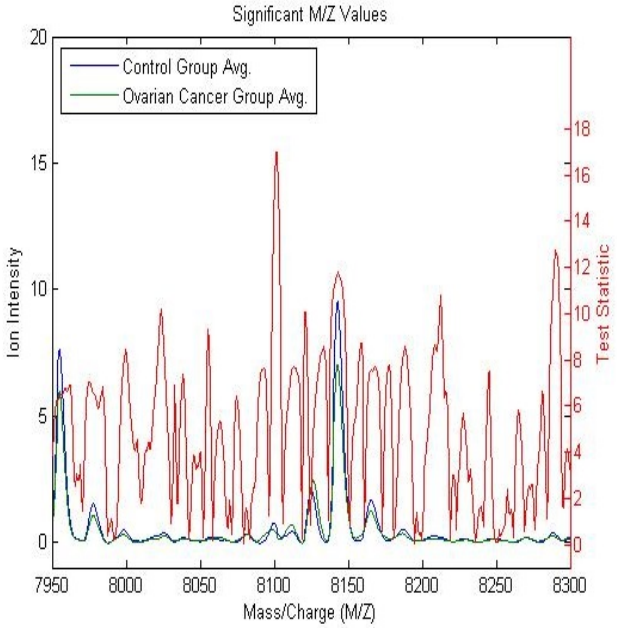

Figure 4: Significant M/Z values

Notice that there are significant regions at high M/Z values but low intensity (~8100 Da.). Other approaches to measure class separability are available in rank-features, such as entropy based, Bhattacharyya, or the area under the empirical receiver operating characteristic (ROC) curve. Now that you have identified some significant features, you can use this information to classify the cancer and normal samples. Due to the small number of samples, you can run a cross-validation using the 20% holdout to have a better estimation of the classifier performance. Features are selected only from the training subset and the validation is performed with the test subset.

After the loop you can assess the performance of the overall blind classification using any of the properties in the CP object, such as the error rate, sensitivity,

ISSN 2348 – 7968

specificity, and others. Table 1 shows some of the classifier.

Table 1 Classifier of detection

Correct Rate	**Error Rate**	**Sensitivity**	**Specificity**	***Prevalence***
0.8186	*0.1814*	*0.8250*	*0.8105*	*0.5581*

This naive approach for feature selection can be improved by eliminating some features based on the regional information. Lilien et al. presented an algorithm to reduce the data dimensionality that uses principal component analysis (PCA), and then LDA is used to classify the groups.

Feature selection can also be reinforced by classification; this approach is usually referred to as a wrapper selection method. Randomized search for feature selection generates random subsets of features and assesses their quality independently with the learning algorithm. Later, it selects a pool of the most frequent good features. Li et al. apply this concept to the analysis of protein expression patterns. The rand-features function allows you to search a subset of features using LDA or a k-nearest neighbour classifier over randomized subsets of features. Also, for better results you should increase the pool size and the stringency of the classifier from the default values in rand-features. The first output from rand-features is an ordered list of indices of MZ values. The first item occurs most frequently in the subsets where good classification was achieved. The second output is the actual counts of the number of times each value was selected. We use histogram to look at this distribution. You will see that most values appear at most once in a selected subset. Figure 5 shows the histogram.

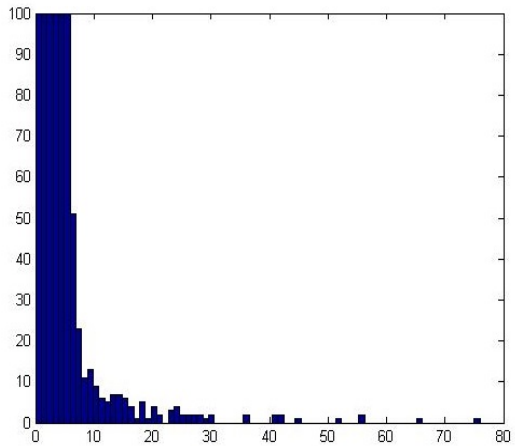

Figure 5: Histogram of selected features

Only a few values were selected more than 10 times. You can visualize these by using a stem plot to show the most frequently selected features.

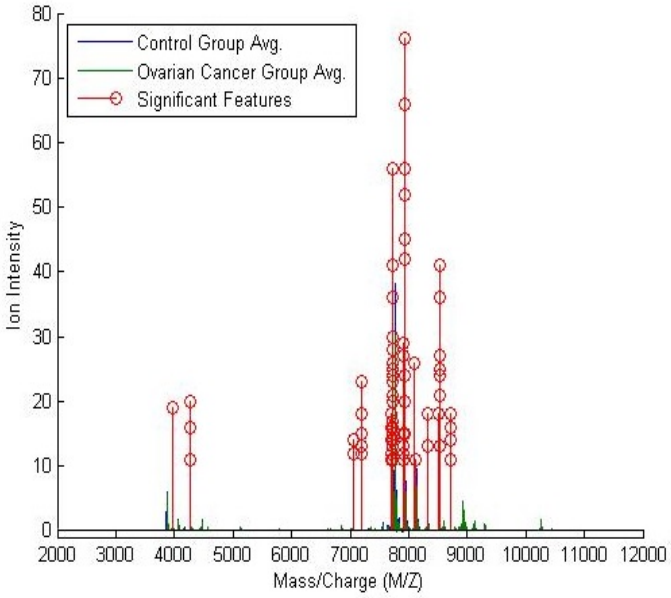

Figure 6: Samples of selected features

These features appear to clump together in several groups. We investigate further how many of the features are significant. The most frequently selected feature is used to classify the data, then the two most frequently selected features are used and so on until all the features that were selected more than 10 times are used. You can then see if adding more features improves the classifier.

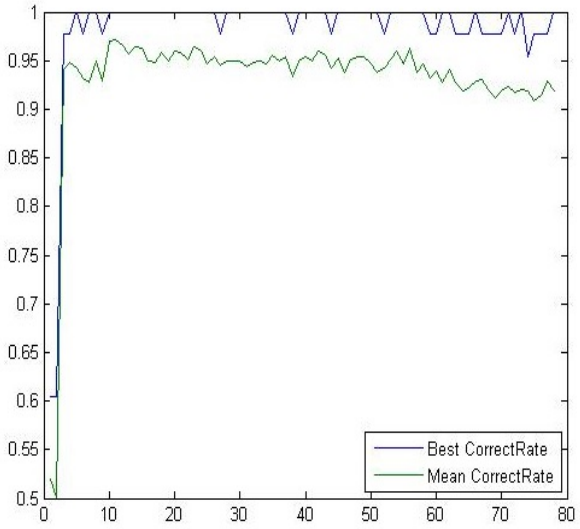

Figure 7: Best Significant Features

ISSN 2348 – 7968

From this graph you can see that for as few as three features it is sometimes possible to get perfect classification. You will also notice that the maximum of the mean correct rate occurs for a small number of features and then gradually decreases. You can now visualize the features that give the best average classification. You can see that these actually correspond to only three peaks in the data. Figure 8 shows the best average classification.

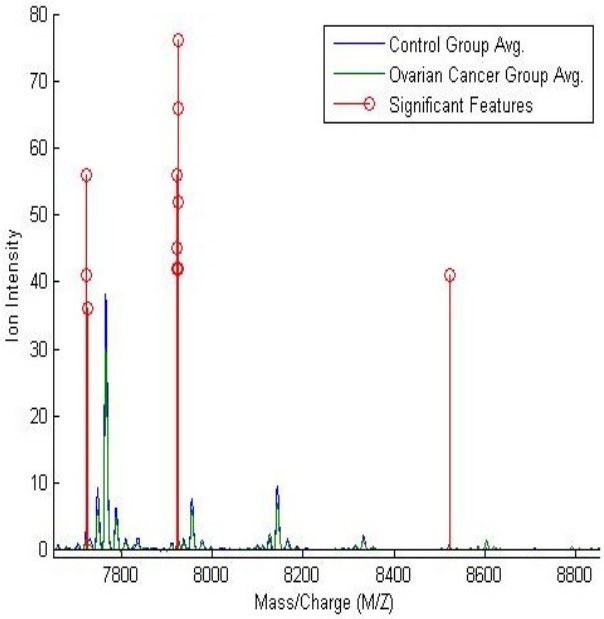

Figure 8: Selected Significant Features

There are many classification tools that you can also use to analyse proteomic data like neural network.

6. Conclusions

This paper gives the overview of ovarian tumor and its detection techniques. The paper explains the idea of how to classify mass spectrometry data and shows some statistical tools that can be used to look for potential disease markers and proteomic pattern diagnostics. Serum proteomic pattern diagnostics can be used to differentiate samples from patients with and without disease. Profile patterns are generated using surface-enhanced laser desorption and ionization (SELDI) protein mass spectrometry. This technology has the potential to improve clinical diagnostics tests for cancer pathologies. The goal is to select a reduced set of measurements or "features" that can be used to distinguish between cancer and control patients. These features will be ion intensity levels at specific mass/charge values.

References

[1] M. Goozner, “Personalizing ovarian cancer screening,” Journal of Natl. Cancer Institute, vol:102, issue:15, pp: 1112–1113, 2010.
[2] J. Tammela and S. Lele, “New modalities in detection of recurrent ovarian cancer,” Curr. Opin. Obstet. Gynecol. 16(1), 5–9 (2004).
[3] V. Nossov et al., “The early detection of ovarian cancer: from traditional methods to proteomics. Can we really do better than serum CA-125,”Am. J. Obstet. Gynecol. 199(3), 215–223 (2008).
[4] S. A. Funt and H. Hedvig, “Ovarian malignancies,” Topics Magn. Res. Imag. 14(4), 329–337 (2003).
[5] Wulfkuhle, J. D., Liotta, L. A., and Petricoin, E. F. (2003) Proteomic applications for the early detection of cancer. Nature Rev. 3, 267–276.
[6] Pusch, W., Flocco M. T., Leung, S.-M., Thiele, H., and Kostrzewa, M. (2003) Mass spectrometry-based clinical proteomics. Pharmacogenomics 4,1–14.
[7] Raj, A. J., Zhang, Z., Rosenzweig, J., Shih, L-M., Pham, T-P., Fung, E. T., Sokoll, L. J., and Chan, D. W. (2002) Proteomic approaches to tumor marker discovery. Arch. Pathol. Lab. Med. 126, 1518–1526.
[8] Menon, U., and Jacobs, I. (2002) Screening for ovarian cancer. Best Pract. Res. Clin. Obstet. Gynaecol. 16, 469–482.
[9] Tuan Zea Tan, Chai Queka, Geok See Ng, Khalil Razvi, “Ovarian cancer diagnosis with complementary learning fuzzy neural network”, Artificial Intelligence in Medicine (2008) 43, 207—222.
[10] Hong Tang, Yelena Mukomel, Eugene Fink, “Diagnosis of Ovarian Cancer based on Mass Spectra of Blood Samples”, 2004 IEEE conference pp: 110-115.