Journal of Molecular Genetics

MEDWELL PUBLICATIONS

Proteomics Approaches for the Early Detection and Diagnosis of Cancer

[1]Saman Hosseini and [2]Mir Behrad Khamesee
[1]*Department of Mechanical Engineering, University of Toronto, Toronto, ON, Canada*
[2]*Department of Mechanical and Mechatronics Engineering, University of Waterloo, Waterloo, ON, Canada*

Key words: Proteomics, cancer, early detection, biomarkers, protein

Corresponding Author:
Saman Hosseini
Department of Mechanical Engineering, University of Toronto, Toronto, ON, Canada

Page No.: 46-56
Volume: 13, Issue 2, 2021
ISSN: 2070-4267
Journal of Molecular Genetics

Abstract: Thousands of patients are annually diagnosed with cancer in Iran and around the globe. The early detection and diagnosis of cancer, the tumorigenesis is in its early phase is crucial for its ultimate control and prevention. Today, the number of people who die from cancer, compared to three decades ago, have declined. Identifying the nature and development of cancer have always attracted a lot of attention from clinicians and scientists. Understanding the cellular changes alone in living species requires basic molecular investigations. Recent advances in molecular biology have been of great help researcher's understanding the complex interaction of genetic alteration, transcription and translation of human cancer. Proteomics studies could play an important role in prevention, early detection and treatment of cancer. Early stage detection is the key to obtain a better outcome for therapeutic intervention of cancer. Although, advances in early stage detection of cancer have come of great help to cancer treatment, most routine screening and diagnosis tools lack sufficient sensitivity and specificity of molecular approaches such as proteomics. With the proteomic technologies emerging, classification and identification of body fluid proteins have been a major focus of scientists. Proteomic analyses have opened a new horizon in screening changes happening in cellular processes to become cancerous, however, it is yet to be perfected using complementary approaches for more accurate diagnosis of cancers. A combination of proteomics approaches like Ciphergen Protein Chip Arrays and SELDI-TOF MS with bioinformatics tools was proved to be effective in the discovery of new biomarkers which further helps the early-stage detection and diagnosis of cancer.

INTRODUCTION

Cancer have always been one of the major challenges in human societies. Cancer remains one of the most common cause of death, accounting for 20% of deaths in the US and 35-100 annual fatal patients out of roughly 100000 cases, globally[1]. Cancer is usually caused by malfunction, this malfunction is caused from genetic

damages due to chemical substances, hormones and sometimes viruses. Therefore, a cancer starts when mechanisms responsible for stabilizing the cell growth process are interrupted[2].

Mutations in two classes of cells, namely tumor suppressor genes and proto-oncogenes, cause cancer. Proto-oncogenes are normal genes that help to regulate the cell growth, however, mutation causes over-stimulation of this process. Tumor suppressor genes inhibit cell division and their mutation leads to excessive cell division. The third type of genes are the caretaker genes that play an important role in cancer. This type of genes encode products that stabilize the genome and their mutation lead to genomic instability[3].

Proteomics: Proteomics is the large-scale study of proteome which is the entire set of proteins expressed by a genome. Proteins control the phenotype of an organism and form a critical part of a living organism which plays important functions in physiologic metabolic pathways of cell[4]. It is estimated that roughly 30,000 genes are responsible for synthesis of the proteome comprised of >500,000 proteins. Most variations are related to consecutive alterations and post-translational changes. Defective proteins are the main reason of cancer and therefor are important indicators for cancer detection and treatment. Moreover, proteins are the main target as well as foundation for design of most medicines. Therefore, proteomics analysis is very useful in early stage detection and control of cancer[5].

Proteomics can improve understanding of cell and living organisms. This approach especially investigates structure, function and operation of proteome, isoforms, structural changes, changes in post-transcription and translation (phosphorylation and glycosylation), interaction with other proteins and medicines which could be of great help in the analogy of mutant and normal proteins[6]. Increased level of complexity as compared to the genome makes proteomics a far more complex and efficient approach than genomics[7].

Proteomics usually implements electrophoresis technique for separating proteins. In this method, protein separation is carried out based on charge and weight, that is isoelectric and molecular weight which can later on help study the changes in Amino acid expression and sequence, causing the formation of new isoforms or changes after transcription and translation like phosphorylation, glycosylation, conjugation and acetylation. One application of proteomics in the design of new drugs to identify and analyze the cell proteome. Identifying cancer-related proteins could help targeting them with drugs that are designed by computer software[8]. Knowing the exact three-dimensional structure helps designing an effective drug to inhibit its activity. Since, different people have different genetic information, they have different protein expression. Therefore, one of the applications of proteomics is to design more specific drugs for the treatment of each person by determining the proteome of any individual[9].

Biomarker: One of the main objective of proteomics is discovering cancer related biomarkers and design of the medicine. Plasma and urine are best sources of studying this kind of proteins. Plasma proteome regularly changes with the cancer status. These changes can be explained by the increase or decrease in the expression of some proteins. Developments in the field of mass spectrometry and bioinformatics can help identifying new biomarkers in cancer. Today, proteomics is considered one of the best and most complete tools for proteomics analysis of biological systems. Biomarkers are important tools for cancer detection and monitoring. The first report of implementing proteomics approach applied for cancer detection was on the detection of ovarian cancer[10].

More than two third of patients suffer from this type of cancer. When this cancer begins to develop, symptoms may be vague or not apparent but they become more noticeable as the cancer progresses and when it reaches its final stage the possibility of treatment becomes scarce. But if in the early stage (stage 1) control of cancer, the patient's chance of survival will increase to >5 years. Therefore, application of proteomics approach or such is necessary for early detection of cancer in order to be able to deal with it[11]. Today, advances made in this field help to identify biomarkers in various cancers as well as the design and function of proteins found. Developments in proteomics and genomics have helped identifying of a wide spectrum of biomarkers with high clinical value. Identification of Biomarkers helps determining the stage of the disease the specific treatment for it[12].

Early stage detection of cancer, evaluation of disease progression, treatment with the use of the most effective techniques and also a measurable factor in human population make biomarkers of paramount importance. These biological molecules describe the physiological condition of the individual[13]. In fact, gene mutations, alterations in protein transcription and translation can all potentially serve as cancer biomarkers. Changes in serum proteome happens with cancer progression can be considered as a biomarker of cancer[14]. Not only the analysis of a cancer is not possible with a single biomarker to have sufficient information on it but also due to changes in the level of expression, various proteins could be valuable[15]. Two-dimensional electrophoresis coupled with mass spectrometry have been primary techniques for in the proteomics study of biomarkers.

Although, spectroscopy is an ideal method for identifying biomarkers, other complementary methods such as high-density small, antibody arrays, protein arrays, molecular arrays and Laser Capture Microdissection (LCM) could be very helpful[16].

MS mass spectrometry is a key development for analysis of biological data, particularly cancer[17]. This technique shows high accuracy for the identification of biomarkers. MALDI-TOF-MS is a powerful tool for the analysis of proteins and peptides. This method is able to detect changes even very in small protein. Considering the importance of proteomics in these types of studies, advances in genomics, epigenome, transcriptome, proteomics and metabolomics, along with proteomics have been effective in the process of identifying biomarkers[15]. Investigating best biomarkers in order to evaluate and study the disease is one of the other challenges in identification of biomarkers. Only 12 different types of cancer biomarkers have been approved by FDA and WHO by 2006, therefore, further studies using other techniques are required to verify the distinction between biomarkers and protein in healthy cells, precancerous and malignant[18].

Application of proteomics in cancer studies: Due to the complexity of biological systems (each cell to produce 107 polypeptides) it is very difficult to study and investigate this system; therefore, extensive research, comprehensive comparative studies carried out by several study groups, is required[19]. Among the applications of proteomics to identify biomarkers in early diagnosis of diseases well. In this regard, it is possible to compare these data with the study of changes isoforms, various modifications in the protein molecules are made to realize the cancer. Applications of proteomics include identifying biomarkers in early diagnosis of diseases. Investigating the data with the study of changes in isoforms and different alterations in the protein molecules can help understanding the cancer status. Detection of cancer biomarkers and distinguishing them from isoforms are very important[20]. Before new techniques such as spectrometry and bioinformatics, Admn and Beadle and Tatum method was used to study proteins.

Recently, labeling of ICT isotopes in mass spectrometry have attracted much attention to measure protein content. One of the applications of proteomics is patient monitoring treatment-specific changes in the patient's serum to further use in the treatment and specific drug design[21]. One of the goals in cancer therapy is targeting glycoprotein protein that plays an essential role in metastasis and immune responses. Branches of oligosaccharide chain of glycoprotein in cancerous cells usually increase which, in turn, increase their attachment to sialic acid that is identified by the lectin and interacts with it. These interactions between sialic acid and lectin make the cancer cells capable of metastasis[22].

One of the applications of proteomics is in drug design, that is, different disciplines such as genomics, proteomics, metabolomiscs, bioinformatics, crystallography X-ray, synthesis chemistry, pharmacology, microbiology, biotechnology and molecular medicine are used. The first step in making the medicine is to identify cancer-causing agent which is conducted by techniques such as genomics, proteomics and metabolomiscs. Genomics can identify faulty cancer-causing genes and thus contribute to cancer proteomics analysis of proteins. The biochemical interactions between genes and metabolomiscs can determine the loss of protein's function. Therefore, by understanding the causes of cancer, we will be able to design drugs to treat people. Proteins are cancer causing factors themselves; therefore, they are considered suitable targets for the drug. Using bioinformatics, we can design drugs that can first be tested for the toxicity of biochemical and on animal and then tested on humans[23].

Literature review: Utilizing proteomics techniques roots back to 70's, however, it took until 1997 for it to be known as proteomics[24]. Two-dimensional electrophoresis technique was first used in 1975 by Farrel and Klose; they succeeded to separate 1100 *E. coli* proteins and spread them across a 2-D gel. The introduction of spectroscopy mounted a great revolution in proteomics. Several methods could be used for protein extraction from cell and tissue[25]. Many factors, including sufficient clinical data and appropriate sampling, affect the results of proteomic analysis. If the analysis is suffering from poor sampling, even the most advanced technology of data analysis would failed to analyze them[26]. With the help of PCR technique, necessary amounts of nucleic acid can be obtained to examine a cell, however, in proteomic methods reproduction of protein samples necessary for the analysis is not possible. Therefore, these approaches have been facing limitations regarding sample preparation and sufficient samples should be prepared for the analyses[27].

Proteomic studies of cancer are subject to problems caused from a mixture of cancerous and healthy cells both present in the tumor tissue; study of the tumor requires purified tumor proteins. Body fluids are a suitable source for proteomic analysis because of availability and ease of sampling and are vital in evaluating the tumor and also repeatability of the test[28].

Using the patient's body fluids have facilitated repeated analyses to assess the patient's status. Another

method is to use the serum for cancer detection and monitoring patient. In fact, serum is one of the best biospecimens to deeply investigate and separate proteins. In spite of recent advances in technology, proteome analysis to identify biomarkers remains a difficult task due to proteins such as albumin; scientists have developed a host of techniques such as immuno-subtractionto help the process. Plasma is another source of biomarkers used for proteomics studies. Plasma is a suitable medium for different proteins in the body. HUPO, in association with different laboratories, initiated the Plasma Proteome Project (PPP) in 2002 as a suitable alternative to serum. Urine is another source to study the protein biomarkers for cancer detection as it can identify cancer biomarkers related to urinary tract and other sorts of cancer[29].

MATERIALS AND METHODS

Samples: Prepared serum samples were obtained from the Johns Hopkins Clinical Chemistry serum banks. This study includes a total of 169 specimens. The cancer group consisted of 103 serum samples from breast cancer patients at different clinical stages: stage 0 (n = 4), stage I (n = 38), stage II (n = 37) and stage III (n = 24). Diagnoses were verified pathologically and sampling was conducted before treatment. Age information was not available on six of these patients. The median age of the remaining 97 patients was 56 years (range, 34-87 years). The non-cancer control group included serum from 25 patients with benign breast diseases (BN) and 41 healthy women (HC). Exact age information was not available from 21 healthy women. The median age of the remaining 20 healthy women was 45 years (range, 39-57 years). The median age of the BN group was 48 years (range, 21-78 years). All samples were stored at 80°C until use.

Protein-chip array analysis: In a typical experiment, 20 μL of each serum sample were mixed with 30 μL of a solution containing 8 mol/L urea and 10 g/L CHAPS in phosphate-buffered saline, pH 7.4. The mixture was vortex-mixed at 4°C for 15 min and diluted to 1:40 (5 μL of mixture plus 195 μL of phosphate-buffered saline) in phosphate-buffered saline. Immobilized Metal Affinity Capture Arrays (IMAC3) were activated with 50 mmol/L $NiSO_4$, the procedure was instructed by the manufacturer (Ciphergen). Diluted samples (50 μL) were applied to each spot on the ProteinChip Array by a 96-well bioprocessor (Ciphergen). The samples were allowed to bind at room temperature for 60 min on a stirrer, then the array was washed twice with 100 μL of phosphate-buffered saline for 5 min and two quick rinses with 100 μL of distilled water. After air-drying, 0.5 μL of saturated sinapinic acid, prepared in 500 mL/L acetonitrile-5 mL/L trifluoroacetic acid, was added twice to each spot. Finally, proteins on the chelated metal (bound by histidine, tryptophan, cysteine or phosphorylated amino acids) were detected with the ProteinChip Reader. Data collection was conducted by an average 80 laser shots (intensity of 240 and a detector sensitivity of 8). Reproducibility was also estimated using two a sample from the healthy controls and one from the cancer patients. Each serum sample was spotted on all eight bait surfaces of one IMAC-Ni array in each of the two bioprocessors. The CV was estimated for the selected mass peaks.

Bioinformatics and biostatistics: All spectra, qualified mass peaks (signal-to-noise ratio >5) with mass-to-charge ratios (m/z) between 2000 and 150 000 were automatically detected. Peak cluster completion was conducted using second-pass peak selection (signal-to-noise ratio >2 within 0.3% mass window) and adding estimated peaks. The peak intensities were normalized to the total ion current of m/z between 2000 and 150 000 all of which were performed using ProteinChip Software 3.0 (Ciphergen). The only additional pre-processing step was logarithmic transformation of the peak intensity data. Such a transformation in general reduces the range of intensity data. As a result, the variance of the transformed peak intensity (across multiple samples) is inclined to be less fickle over the entire length of the spectrum.

ProPeak (3Z Informatics) was used as the software package in order to compute and rank the contribution of each peak towards the optimal separation of two diagnostic groups. It implements the linear version of the Unified Maximum Separability Analysis (UMSA) algorithm which was first reported for use in microarray data analysis (14). The key characteristic of the UMSA algorithm is incorporating data distribution information into a structural risk minimization learning algorithm (15). Therefore, identifying a direction along which the two classes of data are best separated would be facilitated. This direction is represented as weighted sum of the original variables. The weight assigned to each variable in this combination measures the contribution of the variable toward the separation of the two classes of data.

ProPeak, currently, offers three UMSA-based analytical modules. First is the Component Analysis Module which demonstrates each specimen as an individual point in a three-dimensional space. The components (axes) are linear combinations of the original spectrum peak intensities. The axes represent the directions along which two pre-specified groups of data will achieve maximum separability. The two groups of data can be observed to investigate their separation in the component space in an interactive three-dimensional

display. The second module of ProPeak, BootStrap Selection, is used to reduce the complexity of the original data set, in case the separation achieved using combinations of all peaks. This module performs multiple runs of UMSA each of which entails randomly leaving out a fixed percentage of the samples from both groups. The mean, the median and the corresponding SD of the ranks from multiple runs are estimated for each peak. The bootstrap-estimated SD of a peak's rank provides the information about the consistency of the peak's ranking across multiple randomly selected subpopulations of the samples. To establish an objective peak selection criterion, in this study the same bootstrap procedure was also applied to a random dataset that, peak by peak, simulates the distribution of the actual data. The minimum of the rank SDs among all peaks in the simulated random data set was used as the cutoff value for rank SD of the actual data to select a subset of peaks that, in addition to being top-ranked in their contribution to the separation of the data groups, also demonstrated a consistency that was less likely attributable to pure chance. Finally, ProPeak implicates the third module to apply a backward stepwise selection procedure to compute a significance score for each peak. The absolute value of the score is based on the peak's contribution to data separation and is in reverse relation to the order in which it is removed from the initial list of peaks. A positive or negative score indicates relatively increased or decreased expression, respectively, of the corresponding mass peak for the diseased group, whereas the absolute value of the score represents its relative importance toward data separation.

Identifying potential biomarkers that can detect breast cancer at early stages, protein profiles of specimens from stage 0-I, requires breast cancer patients to be compared against those of the non-cancer controls. The analysis involved multiple iterations using all three modules in ProPeak to select from the original full set of mass peaks a small panel of peaks that possessed a consistently high degree of significance in the optimal separation between the two selected diagnostic groups.

After selecting small panel of biomarkers, an evaluation of their ability to detect breast cancer was carried out using the set-aside independent test data set of stages II and III cancer patients. Complementary performance of multiple biomarkers was assessed using a composite index derived by multivariate logistic regression based on the entire data set. Descriptive statistics including p-values from two-sample t-tests and ROC curve analysis were provided for the selected individual biomarkers as well as the composite index. To partially overcome the limitation of lacking a full set of independent test data other than those from the late-stage cancer patients, we used the bootstrap procedure (16) to estimate key performance criteria such as the sensitivity and specificity of the composite index. In this procedure, the patient data set was repeatedly divided through random sampling into a training set to derive a composite index through logistic regression and a test set for computing sensitivities and specificities. The results from multiple runs were then aggregated to form the bootstrap estimate of sensitivity and specificity.

RESULTS

Peak detection and data pre-processing: An analysis of serum proteins retained on the IMAC-Ni2+ arrays was performed on a PBS II mass reader. To acquire the high mass was set to 150 kDa, with an optimization range from 5-30 kDa. A mass accuracy of 0.1% was obtained by external calibration using the All-In-1 Protein Standard (Ciphergen).

Between a total of 147 acceptable detected mass peaks (signal-to-noise ratio >5), 61 peaks had m/z values in 2 to 10 kDa, 30 peaks had m/z values between 10 and 20 kDa, 33 peaks were between 20 and 50 kDa and 23 peaks were between 50 and 133 kDa. Peaks with a m/z<2 kDa were mainly ion noise from the matrix and therefore excluded. Peak intensity was normalized to total ion current (2-150 kDa) and logarithmic transformation was applied. The plots in Fig. 1 illustrate the effect of variance reduction and equalization through logarithmic transformation.

Biomarker selection based on early-stage cancer and non-cancer controls: To identify biomarkers with potential for early stage detection of breast cancer, UMSA was performed using the positive group (early-stage cancer, stage 0-I; n = 42) and the negative group of non-cancer controls (HC+BN; n = 66). First the separability between the two groups was tested by UMSA-derived liner combination of all 147 mass peaks. When the entire protein profiles were compared, the early-stage cancer was separable from the non-cancer group. Figure 2 illustrates the early-stage cancer (red) and the non-cancer (green) data in the UMSA three-dimensional space.

Figure 2a, plot of training data: stage 0-I vs. noncancer using UMSA-derived linear combination of all 147 peaks (Fig. 2b), plot of training data: stage 0-I vs. noncancer using the three selected peaks (Fig. 2c), plot of training and test data: stage 0-I (training data) and stage II-III (independent test data) vs. noncancer (training data), using the three selected peaks.

In order to select biomarkers with consistent performance, UMSA was repeatedly applied for a total of 100 runs, each with a 30% leave-out rate, using the ProPeak BootStrap module. The same procedure was also applied to a simulated random data set. The minimal rank

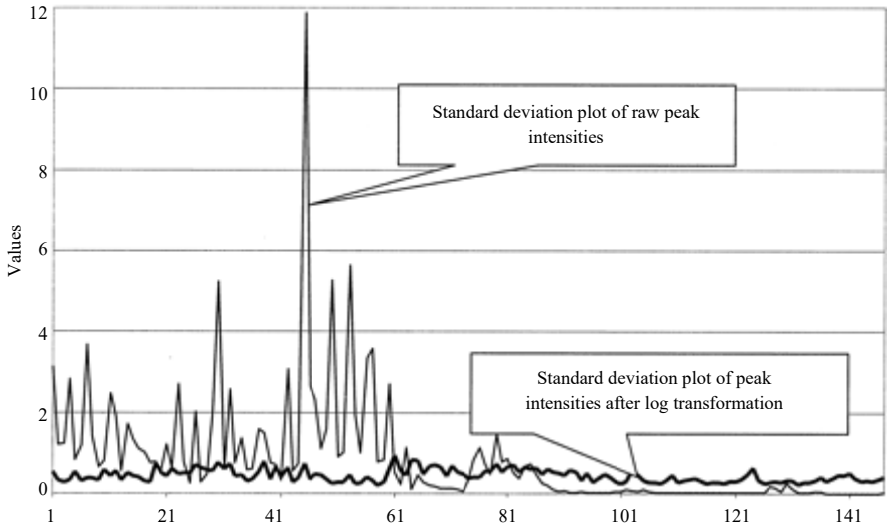

Fig. 1: Effect of logarithmic transformation on data variance reduction and equalization

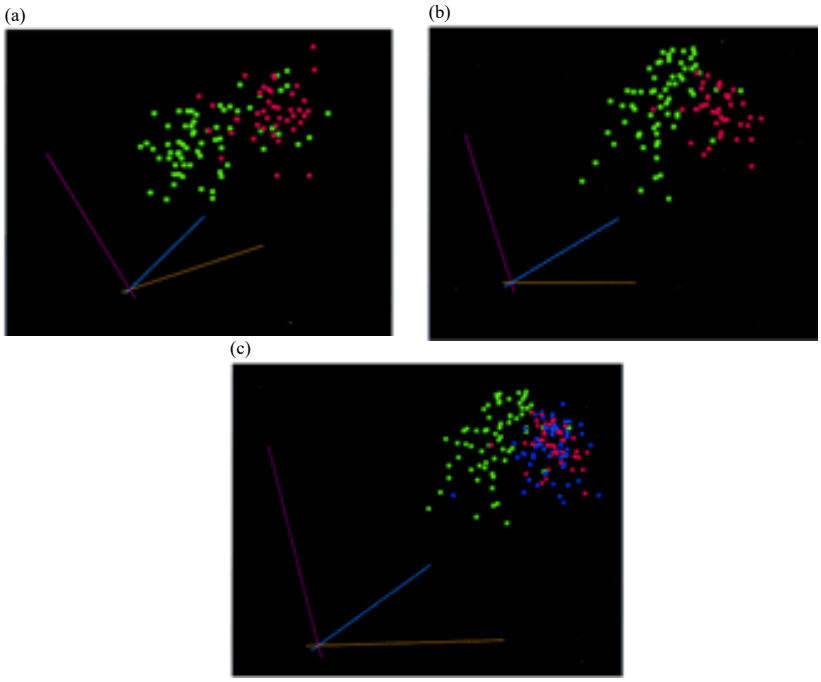

Fig. 2(a-c): Three-dimensional UMSA component plot of stage 0-I (red) or stage II-III (blue) breast cancer vs. non-cancer controls (green)

SD derived from the simulated data was 7.0. Among the peaks with top mean ranks from the actual experimental data, 15 had a rank SD less than this value. They were selected as candidate biomarkers for further analysis. Their mean ranks and the corresponding rank SDs are plotted in Fig. 3.

Horizontal line at 7.0 was the minimum rank SD computed by applying the same procedure to a randomly generated data set that simulated the distribution of the original data. Further ranking of the peaks in this reduced set of candidate biomarkers was performed using the Stepwise Selection module of ProPeak. The absolute value of the relative significance scores of the 15 peaks are presented in a descending order in Fig. 4a which illustrates that the majority of separability between the two groups of data was contributed by the first six peaks.

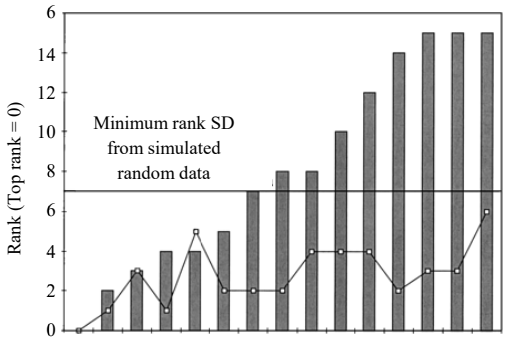

Fig. 3: Fifteen peaks with top mean ranks (▪) and minimal rank SDs (□) derived from ProPeak Bootstrap Analysis

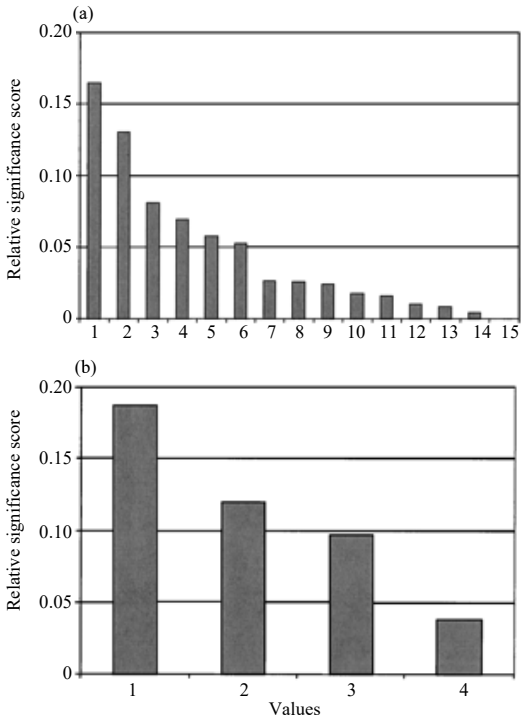

Fig. 4(a, b): Plot of absolute values of the relative significance scores of selected peaks based on contribution toward the separation between stage 0-I breast cancer and the noncancer controls

Among these six peaks, two were identified by ProteinChip Software 3.0 as doubly charged forms of the others. The recognition of both the doubly charged and the singly charged forms of these peaks suggests their importance in disntiguishing the selected two diagnostic groups. By excluding the doubly charged forms, the four unique peaks were further recombined and evaluated using the Backward Stepwise Selection module of ProPeak. The relative significance scores that were recalculated are plotted in Fig. 4b. The top-scored three peaks, designated BC1 (4.3 kDa), BC2 (8.1 kDa) and BC3 (8.9 kDa) were finally selected as the potential biomarkers for detection of breast cancer. Snapshots of three-dimensional plots of stage 0-I or stage 0-III breast cancer against the non-cancer controls using these three biomarkers are shown in Fig. 2, panels B and C, respectively. Between the three biomarkers, BC1 appeared to be down-regulated (scored negative; data not shown) whereas BC2 and BC3 were up-regulated (scored positive; data not shown). This is easily seen in Fig. 5 which illustrates a comparison based on the representative spectra and gel views of the selected biomarkers between cancer and non-cancer controls.

Figure 4a, the 15 peaks selected from ProPeak Bootstrap Analysis with rank SD <7.0. Figure 4b, reevaluated scores of the selected top four peaks.

Figure 5a, BC1 (4.3 kDa), down-regulated in cancer; (Fig. 5b), BC2 (8.1 kDa), up-regulated in cancer; and (Fig. 5c), BC3 (8.9 kDa), up-regulated in cancer. Left panels show the spectrum views; right panels show pseudo-gel views of the same spectra. Both cancer and noncancer representatives were randomly selected with no bias on stages in cancer or between healthy and benign in non-cancer.

Evaluation of the selected biomarkers: The estimated CVs of the log-transformed peak intensities were 6% for BC1, 7% for BC2 and 13% for BC3. The largest CV of 13% belongs to BC3 amongst the three biomarkers. A descriptive statistics of these three biomarkers are summarized in Table 1. Figure 6 shows results of the ROC analysis. BC3 possesses the highest individual diagnostic power [area under the curve (AUC), 0.934] compared with BC1 (AUC, 0.846) and BC2 (AUC, 0.795). Its distributions over the diagnostic groups including clinical stages of cancer patients are plotted in Fig. 7a. Even by considering the sensitivities and specificities of BC3 alone at a cut-off value of 0.8 could differentiate the diagnostic groups which are listed in Table 2 A. The overall sensitivity for breast cancer was 85% and specificity was 91%.

The AUCs, for the composite index for three biomarkers BC1, BC2 and BC3 0.846 are 0.795, 0.934 and 0.972, respectively. Significance for AUC comparisons between individual biomarkers and the composite index is as follows: $p<0.0001$ for BC1 and BC2 vs. the composite index; $p<0.01$ for BC3 vs. the composite index.

Applying a combination of three selected biomarkers: To form a single-value composite index, a multivariate logistic regression was used in order to combine the three selected biomarkers. The descriptive statistics of the

Table 1: Descriptive statistics for BC1, BC2, BC3 and the logistic regression-derived composite index

			Breast cancer patients			
	Noncancer controls (n = 66)		Stage 0-I (n = 42)		Stage II-IIII (n = 61)	
Variables	Mean	SD	Mean	SD	Mean	SD
BC1	0.302	0.312	-0.118	0.244	-0.081	0.258
BC2	0.981	0.358	1.411	0.154	1.295	0.250
BC3	0.526	0.352	0.993	0.193	1.003	0.234
Composite index	-0.375	0.313	0.425	0.257	0.349	0.242

Table 2: Diagnostic performance of BC3 (A) and bootstrap-estimated performance of Logistic Regression (LR)-derived composite index (B)

	Noncancer controls, [1]n			Breast cancer patients by stage, [2]n			
A. BC3							
Cutoff= 0.8	HC[3]	Benign	Subtotal 1	0-I	II[4]	III[4]	Subtotal 1
Positive	0	6	6	37(88%)	29(78%)	22(92%)	88(85%)
Negative	41(100%)	19(76%)	60(91%)	5	8	2	15
Total	41	25	66	42	37	24	103
B. LR-derived composite index[5]	Noncancer controls			Breast cancer patients by stage			
Cutoff= 0	HC	Benign	Subtotal 1	0-I	II	III	Subtotal 1
Positive	-	-	-	93%	85%	94%	93% (85-100%)
Negative	100%	85%	91%(82-100%)	-	-	-	-

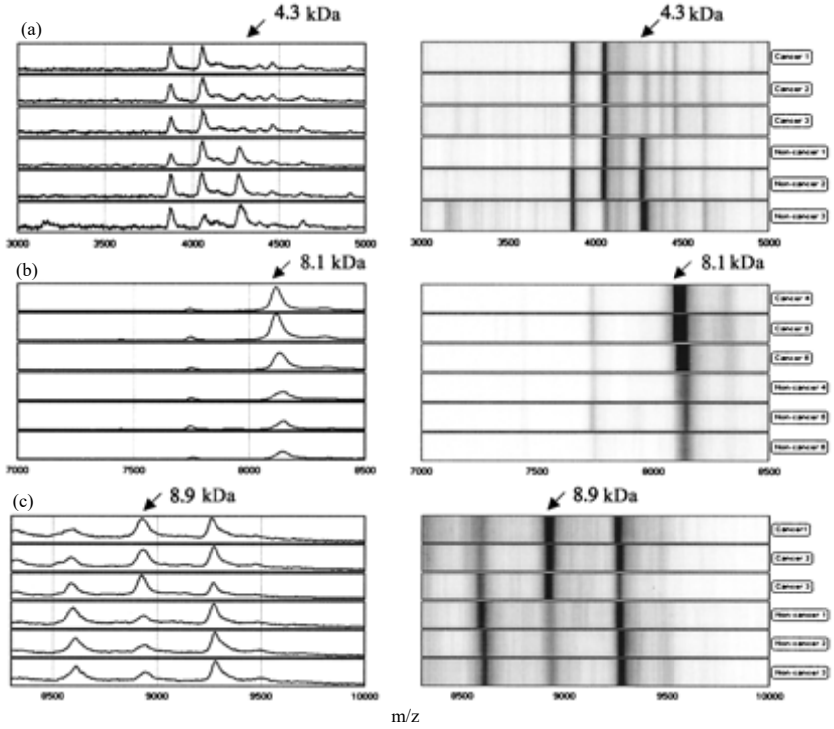

Fig. 5(a-c): Representative spectra and gel views of the selected biomarkers

composite are presented in Table 1. Its distributions over the various diagnostic groups are represented in Fig. 7b. However, ROC curve analysis of the composite index gave a much improved AUC (0.972) compared that of individual biomarkers (Fig. 6).

Estimation of the diagnostic performance of the composite index (20 runs; in each run, 70% samples were randomly selected for composite index derivation and the remaining 30% for testing) was performed using bootstrap cross-validation. The estimated sensitivity (93%) and specificity (91%) are listed in Table 2.

Correlation to tumour size and lymph node metastasis: The three potential biomarkers content were evaluated in relation to pT (tumor size) and pN (lymph node metastasis) categories. No tangible correlation was observed (data not shown).

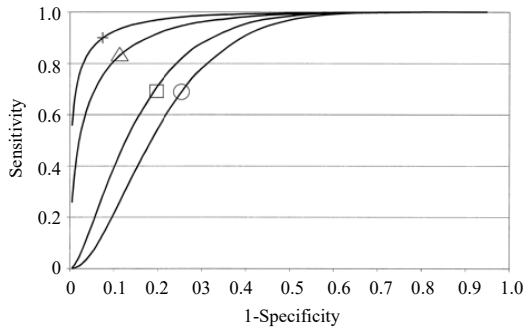

Fig. 6: ROC curve analysis of BC1 (□), BC2 (○), BC3 (△) and logistic regression-derived composite index (+)

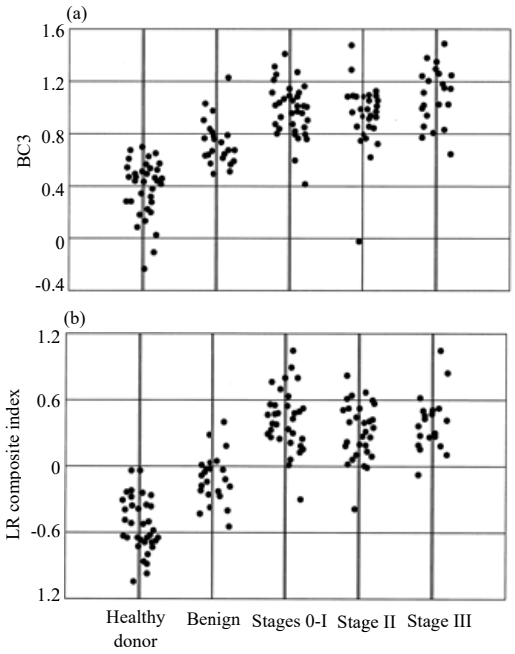

Fig. 7(a, b): Distribution of the selected biomarker(s) across all diagnostic groups including clinical stages of the cancer patients: (a), BC3 alone and (b), Logistic regression-derived (LR) composite index using BC1, BC2 and BC3

DISCUSSION

Proteomics play an important role in prevention, early-diagnosis and treatment with drug design; therefore, it is an invaluable technique for the study of cancer. In most cases detecting cancers is at their malignant state and patients become aware of their situation when malignancy has advanced. Therefore, proteomics by early detection and diagnosis of cancer, even in the premalignant state, helps current or future treatment strategies to have a better chance of cancer treatment. Proteomics gives information on cell processes that control cell division and differentiation in cells, apoptosis in normal cells, cause of abnormalities in healthy cells that initiate the cancer. Methods used in proteomics, however should be improved for better efficiency. Therefore, one of the objectives in perfecting proteomics is to raise the content and purity of extracted proteins and reduce the amount of samples to analyze, automation (i.e., using tools such as robots to replace human in the process), utilizing more efficient software and also using complementary techniques with more sufficient sensitivity and specificity to achieve higher accuracy in detecting the malignancy of cancers[30].

Multifactorial nature of cancer likely necessitates a combination of several markers to effectively detect and diagnose cancer. Such traces of cancer require not only high-throughput genomic or proteomic profiling but also sophisticated bioinformatics tools for complex data analysis and pattern recognition, to be detected.

Simultaneous analysis of the protein profiles of 169 serum samples from patients with/without breast cancer was conducted. The software package ProPeak evaluated each mass peak according to its collective contribution towards the optimal separation of the cancer patients from the non-cancer controls. The two mentioned advances has led to the identification of three discriminatory biomarkers that, if used in combination, achieved both high sensitivity (93%) and high specificity (91%) in early detection of breast cancer patients from the non-cancer controls.

Biomarkers particularly sensitive to differences between early-stage breast cancer patients and non-cancer controls can be found by performing the selection of mass peaks reported here using stage 0-I cancer and non-cancer controls as the training data and more advanced cancer as test data. The biomarkers that were used in the final selection were, however, not sensitive to the stages of cancer patients used in the selection process. In fact, whether the combinations used were stage II vs. non-cancer, stage III vs. non-cancer, or a randomly selected subset of cancer patients at all stages against non-cancer controls, the same three peaks were always selected as the best and most consistently ranked biomarkers. Early detection remains one of the most urgent issues in breast cancer research.

The screening of a large number of potential markers simultaneously can be facilitated through high-throughput profiling of complex protein expression patterns. However, the sample sizes are relatively small for most currently available data sets compared with the total number of detected mass peaks. There is a real danger to mistakenly select mass peaks whose high discriminatory power is purely by chance because of artifacts in the data that are unrelated to the disease process. The use of

high-order nonlinear classification models directly on raw spectrum data may further amplify and mask the influence of such false markers.

In the present study, the UMSA algorithm resulted an efficient model for ranking a large number of peaks collectively according to their contribution to the separation of two predefined diagnostic groups. The ProPeak BootStrap module has brought about random perturbations in multiple runs to examine the consistency of the top-ranked peaks, measured by the SD of computed ranks from multiple runs. To establish an upper cutoff value on a peak's rank SD for its performance not to be considered as purely by chance, the same bootstrap procedure was applied to a randomly generated data set that simulated the distribution of the real data. The minimum value of rank SDs from such "simulated peaks" indicates the degree of consistency that a peak might achieve by random chance. This minimum value was used as the cutoff to help to reduce the original 147 peaks to a subset of 15 peaks for further consideration. The performance of such peaks should be less likely attributable to random artifacts in the data.

The composite index described in this report was derived by simple multivariate logistic regression to achieve simpler results. Further validation of these selected biomarkers, more may use complex and nonlinear classification models to combine the multiple biomarkers. The use of complex modelling methods on carefully screened and tested biomarkers should in general offer a more robust performance than the direct application of such methods on raw data from a large number of mass peaks.

Total specimens analysed in this study to some extent limited the validity of the results. The bootstrap cross-validation estimation of performance brings about statistical confidence on the generalizability of these biomarkers over future data. Further independent validation studies are needed. The specificity of these selected biomarkers for detection of breast cancer needs to be addressed by testing specimens from other types of cancer. Moreover, validation data sets preferably should be from sources different from that of the original training data set. This is one way to ensure that the performance of the selected biomarkers is not influenced by systematic biases between the disease and the control specimens.

The three biomarkers selected showed no significant correlation between their concentrations and the tumour size or lymph node metastasis. Therefore, the discriminatory power of these markers can be attributed to the malignant nature of the tumour rather than its progression. The origin and identity of BC1, BC2 and BC3 are currently under investigation. Furthermore, it is not our intent at this stage to suggest a final diagnostic algorithm based on nonlinear classification.

CONCLUSION

In this study, it was shown that using proteomics approaches such as Ciphergen ProteinChip Arrays and SELDI-TOF MS together with bioinformatics tools could help the discovery of new biomarkers. The panel of three selected biomarkers were used to achieve high sensitivity and specificity for the detection of breast cancer.

REFERENCES

01. Azodi, M.Z., H. Ardestani, E. Dolat, M. Mousavi, S. Fayazfar and A. Shadloo, 2008. Breast cancer: Genetics, risk factors, molecular pathology and treatment. J. Paramed. Sci., Vol. 4.
02. Zali, H., M. Rezaei-Tavirani and M. Azodi, 2011a. Gastric cancer: Prevention, risk factors and treatment. Gastroenterol. Hepatol. Bed. Bench., 4: 11-18.
03. Lodish, H., 2009. Molecular Cell Biology. 6th Edn., Springer, New Jersey, USA., pp: 5-10.
04. Redei, G.P., 2008. Encyclopedia of Genetics, Genomics, Proteomics and Informatics. 3rd Edn., Springer Netherlands, Netherlands, Europe, Pages: 2201.
05. Hayat, M.A., 2009. Methods of Cancer Diagnosis, Therapy and Prognosis: Liver Cancer. 1st Edn., Springer, Netherlands, Europe, Pages: 602.
06. Seyyedi, S.S., M.S. Dadras, M.R. Tavirani, H. Mozdarani, P. Toossi and A.R. Zali, 2007. Proteomic analysis in human fibroblasts by continuous exposure to extremely low-frequency electromagnetic fields. Pak. J. Biol. Sci., 10: 4108-4112.
07. Rozek, W. and P.S. Ciborowski, 2008. Proteomics and Genomics. In: Neuroimmune Pharmacology. Ikezu, T. and H. Gendelman (Eds.), Springer, New Jersey, USA., pp: 725-741.
08. Pooladi, M., S. Sobhi, R.A. Khaghani, M. Hashemi, A. Moradi *et al*., 2013. The investigation of Heat Shock Protein (HSP70) expression change in human brain asterocytoma tumor. Iran. J. Cancer Prevent., 6: 6-11.
09. Gottfries, J., M. Sjogren, B. Holmberg, L. Rosengren, P. Davidsson and K. Blennow, 2004. Proteomics for drug target discovery. Chemom. Intell. Lab. Sys., 73: 47-53.
10. Petricoin, E.F., G.B. Mills, E.C. Kohn and L.A. Liotta, 2002. Proteomic patterns in serum and identification of ovarian cancer. The Lancet, 360: 170-171.
11. Wulfkuhle, J.D., L.A. Liotta and E.F. Petricoin, 2003. Proteomic applications for the early detection of cancer. Nat. Rev. Cancer, 3: 267-275.
12. Ludwig, J.A. and J.N. Weinstein, 2005. Biomarkers in cancer staging, prognosis and treatment selection. Nat. Rev. Cancer, 5: 845-856.

13. Srinivas, P.R., M. Verma, Y. Zhao and S. Srivastava, 2002. Proteomics for cancer biomarker discovery. Clin. Chem., 48: 1160-11696.
14. Srinivas, P.R., B.S. Kramer and S. Srivastava, 2001. Trends in biomarker research for cancer detection. Lancet Oncol., 2: 698-704.
15. Rezaei-Tavirani, M., H. Zali, F.R. Jazii, M.H. Heidari, B. Hoseinzadeh-Salavati, F. Daneshi-Mehr and K. Gilany, 2010. Introducing aldolase C as a differentiation biomarker: A proteomics approach. Arch. Adv. Biosci. Vol. 1, No.1.
16. Bichsel, V.E., L.A. Liotta and E.F Petricoin 3rd, 2001. Cancer proteomics: From biomarker discovery to signal pathway profiling. Cancer J., 7: 69-75.
17. Mechref, Y., Y. Hu, A. Garcia and A. Hussein, 2012. Identifying cancer biomarkers by mass spectrometry-based glycomics. Electrophoresis, 33: 1755-1767.
18. Uen, Y.H., K.Y. Lin, D.P. Sun, C.C. Liao, M.S. Hsieh *et al*., 2013. Comparative proteomics, network analysis and post-translational modification identification reveal differential profiles of plasma Con A-bound glycoprotein biomarkers in gastric cancer. J. Proteom., 83: 197-213.
19. Reymond, M.A., J.C. Sanchez, G.J. Hughes, K. Gunther, J. Riese *et al*., 1997. Standardized characterization of gene expression in human colorectal epithelium by two-dimensional electrophoresis. Electrophoresis, 18: 2842-2848.
20. Kruger, T., J. Lautenschlager, J. Grosskreutz and H. Rhode, 2013. Proteome analysis of body fluids for amyotrophic lateral sclerosis biomarker discovery. Proteomics. Clin. Appl., 7: 123-135.
21. Goldknopf, I.L., 2008. Blood-based proteomics for personalized medicine: Examples from neurodegenerative disease. Proteomics, vol. 5, No. 1. 10.1586/14789450.5.1.1
22. Kim, Y.J. and A. Varki, 1997. Perspectives on the significance of altered glycosylation of glycoproteins in cancer. Glycoconjugate J., 14: 569-576.
23. Cheng, Y., T. LeGall, C.J. Oldfield, J.P. Mueller and Y.Y.J. Van *et al*., 2006. Rational drug design Via. intrinsically disordered protein. Trends Biotech., 24: 435-442.
24. Wilkins, M.R., C. Pasquali, R.D. Appel, K. Ou and O. Golaz *et al*., 1996. From proteins to proteomes: Large scale protein identification by two-dimensional eletrophoresis and amino acid analusis. Nature Biotechnol., 14: 61-65.
25. Pooreydy, B., F. Tajik, M. Jafari, M. Karimi, M. Rezaei-Tavirani *et al*., 2013. Organelle isolation for proteomics: Mitochondria from peripheral blood mononuclear cells. J. Paramed Sci., 4: 78-86.
26. Canas, B., C. Pineiro, E. Calvo, D. Lopez-Ferrer and J.M. Gallardo, 2007. Trends in sample preparation for classical and second generation proteomics. J. Chromatogr. A., 1153: 235-258.
27. Moon, H., A.R. Wheeler, R.L. Garrell and J.A. Loo, 2006. An integrated digital microfluidic chip for multiplexed proteomic sample preparation and analysis by MALDI-MS. Lab. Chip., 6: 1213-1219.
28. Zali, H., G. Ahmadi, R. Bakhshandeh and M. Rezaei-Tavirani, 2011b. Proteomic analysis of gene expression during human esophagus cancer. J. Paramed Sci., 2: 2008-4978.
29. Albalat, A., J. Franke, J. Gonzalez, H. Mischak and P. Zurbig, 2012. Urinary proteomics based on capillary electrophoresis coupled to mass spectrometry in kidney disease. Clin. Appl. Capillary Electrophoresis., 919: 203-213.
30. Shi, T., D. Su, T. Liu, K. Tang, D.G. Camp *et al*., 2012. Advancing the sensitivity of selected reaction monitoring-based targeted quantitative proteomics. Proteomics, 12: 1074-1092.