Biomarker Discovery through the use of Proteomics Tools

by

Rama Yakubu
Clarkson University

Biomarker Discovery through the use of Proteomics Tools

A Thesis Proposal by

Rama Yakubu

Biomolecular Science & Chemistry

Mentor: Dr. Costel C. Darie

March 12, 2010
Abstract

Serum biomarkers are proteins or peptides that either appear in or disappear from the sera of people with a particular disease, compared with serum from normal subjects. The area of proteomics, which is the study of protein structure and function, has provided useful tools for the detection of such serum biomarkers. The discovery of new serum biomarkers could help to establish an early warning system, which could decrease the deaths due to cancer through prognosis and proper diagnosis in early stages of the disease. We propose to search for protein biomarkers in the human serum samples of cancerous patients using proteomic research techniques with subsequent use of mass spectrometry for peptide identification. Novel to our approach of biomarker discovery is the analysis of serum biomarkers in SDS-PAGE non-reducing conditions, which leaves existing disulfide bonds between cysteine residues of the same or other proteins intact.

Introduction

Every year a large percentage of American deaths (23.1%) are due to cancer, thus we have realized the need for an improved method of early identification of protein serum biomarkers in patients [1]. Serum biomarkers are proteins or peptides, indicative of certain diseases through their appearance or disappear from the sera of people affected by a particular disease. Compared to the serum of normal subjects a difference in the abundance of these biomarkers should become apparent. Successful biomarkers for diagnosis and prognosis have been identified in malignant tumors through standard proteomics approaches. These biomarkers, if discovered, could aid in establishing an early warning system, which could decrease the deaths due to cancer through prognosis and proper diagnosis in early stages of the disease. Such biomarkers have already been detected in malignant tumors from breast, prostate, pancreas, colon and ovarian cancer.

The search for cancer-specific serum biomarkers is crucial in ovarian cancer as it has few early symptoms and is often diagnosed in an advanced stage. This leads to low survival rate due to poor prognosis of the disease. It is the most common cause of death from gynecological cancer [2]. Breast cancer is the most prominent form of cancer in women with 1.15 million new cases in 2002 and 41000 deaths due to breast cancer in the same year [3]. These statistics present a strong driving force for the search for improved biomarkers for breast cancer. One key assumption in the discovery of these biomarkers is that blood, which continuously passes through cancerous tissue will signify the presence of disease through the differential expression of specific proteins in the blood, which may be identified as biomarkers. These proteins are proposed to carry important metabolic or clinical information, which may aid in early detection of disease and monitoring of treatment.

Currently a study on the discovery of biomarkers for breast cancer along with others are implementing 2-D-differential gel electrophoresis (2-D DIGE) and MALDI-TOF MS to investigate protein expression alterations in breast cancer serum as compared to control serum samples. Already four up-regulated spots in patient groups have been identified as three proteins: proapolipoprotein A-I, transferrin and hemoglobin. Three down –regulated spots in patient groups were identified as the proteins apolipoprotein A-I, apolipoprotein C-III and haptoglobin α2. Ultimately three candidates were selected as potential biomarkers for validation by immunochemical methods in all individual serum samples: transferrin, haptoglobin α2 and apolipoprotein A-I [3].

Adenocarcinoma of the pancreas (PA), being the fourth most common cause of cancer deaths in men and women in the United States with a five year survival rate of only 4%, has one of the lowest survival rates of all cancers. Its high mortality rate is attributed to the commonly advanced stage of the cancer at its point of diagnosis and treatment. One reason for the late detection of the disease is its asymptomatic initiation and progression. Resection, a highly invasive procedure is the only currently known treatment for this disease and is only effective in 10 – 15% of patients with localized disease [4].

There are already some serum biomarkers in use, which include CA 125, CEA, CA 15.3 and TPS. However, a common weakness of these biomarkers is low specificity and low detection sensitivity. Note that the combined use of two or more known biomarkers does not improve sensitivity, rather this
combination decreases specificity. Therefore it is urgently necessary to discover serum biomarkers that are reliable in terms of specificity and have a relatively high sensitivity. With these requirements it is not surprising that the number of new biomarkers reaching clinical use is exceptionally low [3].

Although a number of serum biomarkers have been identified, cancer antigen 125 (CA 125) is one of the few carefully investigated biomarkers for ovarian cancer screening. This biomarker is currently being used in combination with transvaginal ultrasound in order to screen for and improve sensitivity of ovarian cancer detection. Despite the elevated CA 125 levels in 83% of patients with epithelial ovarian cancer, 50% of patients with stage 1 ovarian cancer and 90% of patients with more advanced stages, more sensitive biomarkers or pattern of markers are required [5].

Background

Mass spectrometry (MS) is a method that can analyze a variety of chemical and biological compounds, from products of chemical synthesis or degradation to biological building blocks and their products, including proteins, nucleic acids, lipids, and glycans. Within the mass spectrometry field different specialized fields have emerged. They have been named based on the chemical or biological entities that are analyzed, including: proteomics, lipidomics, glycomics and others. Proteomics analyzes proteins, peptides, aglycans, protein interactions, or post-translational modifications (PTMs). Proteomic analysis is usually qualitative for protein identification and quantitative for estimation of the amounts of proteins. These analyses are usually performed using Matrix Assisted Laser Desorption Ionization Mass Spectrometry (MALDI-MS) and Electrospray Ionization Mass Spectrometry (ESI-MS) mass spectrometers.

The workflow in a proteomics experiment involves sample fractionation by 1, 2, n biochemical approaches, followed by enzymatic digestion (usually trypsin), peptide extraction, and MS analysis [6]. When the peptide mixture is analyzed by MALDI-MS, the proteins of interest are identified using a procedure named peptide mass fingerprinting. Alternatively, the peptide mixture is further fractionated by HPLC on different columns (usually reverse phase HPLC or LC), followed by ESI-MS analysis. The combination of LC and ESI-MS is usually named LC-MS/MS, and analysis of a protein using this approach provides not only the protein identity, but also sequence information for that particular protein. In addition to qualitative information provided by MALDI-MS or LC-MS/MS analysis, MS may also provide quantitative information about a particular protein. Methods such as differential gel electrophoresis (DIGE) [7], isotope-coded affinity tag (ICAT) [8], stable isotope labeling by amino acids in cell culture (SILAC) [10], absolute quantitation (AQUA) [9], multiple reaction monitoring (MRM) [11], or spectral counting [12] allow detection, identification and quantification of proteins or peptides. These methods are currently known as functional proteomics and are widely used in basic research. Some of these methods can be specifically applied to the identification and quantitation of tissue and serum biomarkers.

Serum biomarkers are proteins or peptides that either appear in or disappear from the sera of people with a particular disease, compared with serum from normal subjects. Successful biomarkers for diagnosis and prognosis have been identified for a wide variety of diseases, from malignant tumors—such as breast, prostate, ovarian, colon cancer—to benign tumors. Identification of biomarkers involves employment of the standard proteomics approaches mentioned above and can include both MALDI-MS and LC-MS/MS analysis of proteins or peptides. However, although much effort is invested in this field, these approaches also have limitations and there is always room for improvement. Therefore, any new approaches for identification and quantification of the serum biomarkers with better, faster, and cheaper methods would be helpful in reducing the number of people with diseases though their diagnosis in early disease phases.

In our current and future research we are using a novel approach for biomarker discovery. The goal of our proposed research is three-fold: 1) to establish a new method for the analysis of serum biomarkers, 2) to identify new serum biomarkers and 3) to validate them. Initially in the experimental workflow the serum samples will be separated on SDS-PAGE NR. Each gel will separate serum samples
from patients with a specific type of cancer and sera from normal patients. We have serum samples from normal patients and from patients with breast cancer, cervical cancer, prostate cancer and prostate cancer post-prostatectomy, as well as various types of cancer in both males and females. Potential protein biomarkers will be excised from the gel, digested with trypsin and further analyzed by LC-MS/MS. The data will then be submitted to the MASCOT database for protein identification. **We hypothesize that by separating the serum samples under non-reducing conditions, the disulfide-linked proteins will be easily identified by mass spectrometry, avoiding the additional steps in sample preparation normally used in standard proteomics preparation.** We also hypothesize that this approach will be selective: only the extracellular/secreted proteins or proteins with an extracellular domain will be identified. Although easy and simple, this approach has never been tested.

Based on the preliminary research our research design that we have proposed is doable and realistic. The procedural steps in our proteomics experiments are well known and will serve to characterize the disulfide-linked proteins that are different in the serum samples of normal patients and cancerous patients. There is extensive experience in the characterization of proteins as reflected in the preliminary data and my advisor’s publication record [13].

**Research Methodology**

Currently, we are investigating several biomarker discovery approaches such as the analysis of serum biomarkers using sodium dodecyl sulfate polyacrylamide gel electrophoresis in non-reducing conditions (SDS-PAGE NR). This method can identify proteins with cysteine residues interconnected within the same protein or between two or more proteins known as disulfide bonds. Novel to our approach of biomarker discovery is the analysis of serum biomarkers in SDS-PAGE non-reducing conditions, which leaves existing disulfide bonds between cysteine residues of the same or other proteins intact. These disulfide-linked proteins may be identified by comparing the serum samples from normal patients to those of diseased patients, thus identifying potential biomarkers.

The most commonly used tools for serum proteomic pattern analysis is 2-D electrophoresis coupled with MS. This combination not only facilitates the identification of differential protein expression but also is a useful tool in investigating chemically modified protein isoforms that have undergone post-translational modification. These modifications are almost universal among proteins and are needed in order to activate the protein. The purpose for the use of this technique is the detection of low-abundance serum proteins with differential expression in order to identify potential protein biomarkers. The utilization of this technique is preceded by the depletion of high abundance serum proteins.

Our standard proteomic approach involves initially fractionating the sample by SDS-PAGE-NR, followed by excision of the gel piece and enzymatic (trypsin) digestion. The next step is peptide extraction and analysis through mass spectrometry for protein identification. Through the combination of LC and MS, also known as LC-MS/MS, we are able to determine the identity of the protein along with its amino acid sequence information. This qualitative information is supplemented by quantitative information about the protein, which may allow us to detect not only the occurrence or disappearance of a serum protein but also its relative amounts as compared to other serum proteins. This will allow us to detect differentially expressed proteins in order to create a more complex protein profile, which may give us more information about the condition of the patient on an intra- and extracellular level.

SDS-PAGE NR will be a self-made, 4-13% acrylamide bisacrylamide and will be performed in a Protean XI gel electrophoresis system (Bio-Rad). The samples will be combined with SDS-PAGE sample buffer, except that it will not contain the reducing agent (DTT or beta-mercapto-ethanol. For the second dimension SDS-PAGE, the gel lanes will cut and reduced for 1 hour in 2% (v/v) beta-mercaptoethanol under moderate shaking, washed and then separated in the second SDS-PAGE dimension. The gel bands will be excised and washed in HPLC water, cut into small pieces and incubated with 50% acetonitrile (ACN) in digestion buffer (25 mM ammonium bicarbonate, pH 8.0) for 20 min and then in 100% ACN under moderate shaking, followed by drying in Speed-vac. This procedure was repeated
three times. After drying, gel pieces were rehydrated in digestion buffer containing 5 mM DTT and incubated for 1 h at 56°C. Reduced cysteine residues in the proteins will be blocked by replacing the DTT solution with 55 mM iodoacetamide in digestion buffer, for 45 min at room temperature with occasional vortexing. Gel pieces will be dehydrated, dried, and rehydrated twice. Dried gel pieces will then be digested overnight at 37°C in digestion buffer containing 15 ng/µl trypsin. The resulting peptides will be extracted twice with 5% formic acid/50 mM ammonium bicarbonate/50% ACN and once with 100% ACN under moderate shaking. Peptide mixture will then be dried in a Speed-vac, solubilized in 20 µl of 0.1% formic acid/2% ACN and cleaned with a P10 ZipTip µ-C18 (Millipore Corporation, Billerica, MA) and dried again. For HPLC analysis, the samples will be resuspended in 10 µl of 0.1% formic acid/2% ACN.

**LC-MS/MS analysis and protein identification and quantification.** The resulting peptide mixture will be analyzed by reverse phase liquid chromatography and mass spectrometry (LC-MS/MS) using a Waters Alliance 2695 HPLC coupled to a Q-Tof Micro mass spectrometer (Waters, Milford, MA). The peptides will be loaded onto a 100 µm x 10 mm nanoAquity BEH130 C18 1.7 µm HPLC column (Waters, Milford, MA) and eluted over a 120 minutes gradient of 10-55% acetonitrile in 0.1% formic acid at a flow rate of 0.250 µl/min. The column will be coupled to a Picotip Emmitter Silicatip nano-electrospray needle (New Objective, Woburn, MA). MS data acquisition will involve survey MS scans and automatic data dependent MS/MS of 2+, 3+ or 4+ ions. The MS/MS will be triggered when the MS signal intensity exceeded 10 counts/second. In survey MS scans, the three most intense peaks will be selected for CID and fragmented until the total MS/MS ion counts will reach 10,000 or for up to 6 seconds each. The raw data will processed using ProteinLynx software and the resulting pkl files will be submitted for database search and protein identification to the open source Mascot search engine (Matrix Science, London, UK, www.matrixscience.com). The Mascot search will provide a list of proteins for each gel band. Only proteins identified by two or more peptides and with a Mascot score >50 will be considered. We have been able to obtain interpretable MS/MS spectra for low femtomolar amounts of standard peptides.

**Preliminary Studies:**

![Figure 1: SDS-PAGE of sera from normal patients (patient # 81) and patients with various cancers (patients # 1 & 2, from left to the right), under non-reducing (-DTT) and reducing (+DTT) conditions. The molecular weight marker is shown.](image)

Scientists analyze many proteins from a variety of sources as potential biomarkers for the diagnosis and prognosis of diseases. Currently used analytical technologies include 2D PAGE, multi-dimensional chromatography, MS, antibody microarrays, micro-fluidics, label-free technologies and many others. Improving the speed, quality, and/or costs of current approaches would advance biomarker discovery.

We are currently investigating several biomarker discovery approaches. One such approach involves the analysis of serum biomarkers using SDS-PAGE NR. This method can identify proteins with cysteine residues interconnected within the same protein or between two or more proteins. These disulfide-linked proteins may be identified by comparing the serum samples from normal patients and patients with a disease, and are potential biomarkers. No other researcher or research group used or uses this approach.

In Figure 1 the separation of serum samples from normal and cancer patients are shown on SDS-PAGE under non-reducing (NR) and reducing (R) conditions. In reducing conditions SDS-PAGE R no difference is observed in the two types of samples. However, significant differences between normal and
cancer serum samples were observed in non reducing conditions on SDS-PAGE NR. Apart from the expected Immunoglobulin G (IgG) band at about 150 kDa in all samples, a band around 110 kDa in serum samples from patients with cancer was detected, but not in the serum from the normal patient. Note the greater intensity of the band in patient # 2. Since the 110 kDa band was observed only in SDS-PAGE NR, this suggests that a minimum of two different disulfide-linked proteins are present in this band.

Intra-disulfide bridges can be excluded because we did not see a similar band in SDS-PAGE R. To identify the proteins responsible for this band, the gel pieces were excised from both normal (sample # 81) and cancer (samples # 2) sera, subsequently the gel pieces were digested with trypsin, extracted and analyzed by LC-MS/MS. Mascot database search identified the 110 kDa band as haptoglobin (45 kDa), a well known marker for cancer and inflammation [14]. Haptoglobin is an intensely processed protein that splits into alpha and beta subunits. These two subunits form a hetero-tetramer that contain two alpha and two beta subunits, all disulfide-linked. Therefore, haptoglobin identified by LC-MS/MS is indeed a multi-subunit, disulfide-linked protein that could be used as a diagnostic marker for inflammation. Upon reviewing patient # 2 file, it was discovered that in addition to cancer, the patient also had a stomach inflammation (gastritis). To test whether the 110 kDa band could be a potential biomarker for inflammation, we extended our studies and analyzed samples from 25 patients with different types of cancer using SDS-PAGE NR. Only one additional sample (sample # 4) had a band similar to the band observed in sample # 2 (Figure 1) and LC-MS/MS analysis identified it as haptoglobin. Again, when we analyzed the file of patient # 4, we observed that this patient, in addition to having cancer, also suffered from an inflammatory process.

Since haptoglobin is a well-known marker for acute infections and inflammation—with its serum levels increased at least ten times in acute infections/inflammation compared with levels from normal patients—we sought to compare the levels of haptoglobin in the sera of patients with cancer and inflammatory processes (samples # 2 & # 4) and normal (sample # 81) patients. To do this, we used both Mascot score and the relative intensity of the peaks that corresponded to haptoglobin peptides identified in both normal and cancer samples. Mascot score reflects the probability with which a protein may be identified in a database search, by comparing the mass spectrometric experimental data with the theoretical data obtained from in silico digestion of the proteins from that database. On other words, the higher the Mascot score for a protein, the higher the probability that the identified protein is real. The Mascot score by itself is a relative measure of quantitation for a particular protein. An increased Mascot score in the bands from cancer samples, compared with control non-cancerous samples, reflects a higher amount of that protein in the cancer samples. In our experiments, haptoglobin was identified in the cancer samples # 2 (Mascot score 645) and # 4 (Mascot score 345), as well as in the normal sample # 81 (Mascot score 70). Therefore, based on the Mascot score, the cancer sera contained at least five to ten times more haptoglobin, compared with the normal serum. To confirm these findings, we compared the intensities of two different peptides identified in both normal and cancer sera. Such an analysis is shown in Figure 2.

**Figure 2:** Label-free, semi-quantitative analysis of haptoglobin. Comparison of the intensities of MS spectra for peaks corresponding to peptide TEGDGVYTLNDKK that are part of haptoglobin. This peptide was identified by SDS-PAGE NR and LC-MS/MS in all samples analyzed (#2, #4 and #81). Samples #2 and #4 were from patients with cancer (and inflammation) and sample # 81 was from a normal patient. The intensity scale for the spectra for each individual peptide was identical.
One double-charged peak with m/z of 720.29 in sample #2 and 720.31 in sample #4 had a much higher intensity than the peak with m/z of 720.30 from sample #81 (Figure 2). Similarly, one triple-charged peak with m/z of 729.97 in sample #2 and 730.00 in sample #4 had a much higher intensity than the peak with m/z of 729.99 from sample #81 (data not shown). These data suggest that comparison of the Mascot scores and intensities of the MS spectra are a good measure of the relative abundance of a protein in a label-free approach. These data also suggest that samples #2 and #4 contained more haptoglobin than sample #81, confirming the medical records that patients #2 and #4, in addition to having cancer, also suffered from an inflammatory process.

We also tested our approach in a blind experiment (without knowing where the samples were collected from) and accurately identified that a normal patient (#89), with a high level of serum haptoglobin (as determined by SDS-PAGE NR) had a prostate inflammation (data not shown). Therefore, our approach is very accurate for detecting an inflammatory process without prior knowledge about the medical status of a patient. We have demonstrated that we are able to find new approaches for identification and validation of biomarkers and are confident that we will succeed to accomplish our proposed work. An example of our current work is reflected in Figure 3, where various serum samples were separated by SDS-PAGE NR and several potential serum biomarkers (in the 110, 130 and 180 kDa range) may be observed.

**Figure 3:** 4-13% gradient SDS-PAGE NR of serum samples from normal males (NM, #15 - #19), and samples from people with prostate cancer (PC, #1 - #7), various cancer (CM, cancer male, #8 - #13), rheumatoid arthritis (#14).
# Timeline:

<table>
<thead>
<tr>
<th>Month</th>
<th>Stage</th>
<th>Goals of Each Stage</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 2010</td>
<td>Initial Research</td>
<td>● Complete Mass Spectrometry Training</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Run 1D PAGE serum samples in non-reducing conditions; prostate cancer, prostate</td>
</tr>
<tr>
<td></td>
<td></td>
<td>cancer post prostatectomy &amp; normal</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Run 2D SDS-PAGE (from 1D)</td>
</tr>
<tr>
<td>April 2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>August 2010</td>
<td>Protein Purification / Growth</td>
<td>● 2D SDS-PAGE stain (Coomassie, Silver), analysis, complete the optimization of</td>
</tr>
<tr>
<td></td>
<td>Factor Analysis</td>
<td>the sample preparation,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● complete the 2D PAGE analysis of undepleted serum samples,</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● start 2D SDS-PAGE of depleted serum samples, staining and analysis</td>
</tr>
<tr>
<td>September 2010</td>
<td></td>
<td>● Excision of target proteins from bands and analysis through LC-MS/MS</td>
</tr>
<tr>
<td>October 2010</td>
<td></td>
<td>● Complete data analysis and verification/validation and quantification through label-</td>
</tr>
<tr>
<td></td>
<td></td>
<td>free MS Western Blotting analyses</td>
</tr>
<tr>
<td>November 2010</td>
<td></td>
<td>● Complete Rough Draft of Thesis</td>
</tr>
<tr>
<td>December 2010</td>
<td></td>
<td></td>
</tr>
<tr>
<td>January 2011</td>
<td>Final Results</td>
<td>● Correct Discrepancies between Data Sets</td>
</tr>
<tr>
<td></td>
<td></td>
<td>● Finish Thesis Paper</td>
</tr>
<tr>
<td>February 2011</td>
<td>Thesis Completion</td>
<td>● Turn In Thesis Paper and Present Research</td>
</tr>
<tr>
<td>March 2011</td>
<td></td>
<td></td>
</tr>
<tr>
<td>April 2011</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
References


