Biomarker detection with the help of proteomic approaches

Pratik Mehta*1, Keyur Patel1, Ujjwal Sahoo1, A.K.Sen1, Dhanya B1 and Sachinandan Basak2

1Department of Pharmacy, Sumandeep Vidyapeeth, Vadodara-391760, Gujarat, India.
2Dr. Reddy's Laboratories Ltd, Integrated Product Development, Bahupally, Qutubullapur, R.R.Dist, 500090, India.

*Corres. author: mpratik.10@gmail.com

Abstract: The recent advancements in the mass spectrometry technologies have brought the ability to gather mammoth information of data characterizing the proteomes of complex mixtures. Also these advancements lead to influence the data-gathering capability to conduct comparative analysis of biofluids from healthy as well as disease-affected patients for the identification of highly specific biomarkers along with the development of mass spectrometry-based diagnostic platforms. The main focus is into optimizing the biofluid proteome coverage, which can be obtained using these technologies, leaving proteomics on the edge to make an important impact in disease diagnostics in the future.

Keywords: biomarkers, serum, mass spectrometry, proteomics.

Introduction:
The swift rise in the focus on proteomics technologies has resulted in an exceptional ability to identify a very huge number of proteins in complex biological mixtures. While the development of this technology was stuck to the analysis of simple prokaryotes1 and eukaryotes2, much of the attention has turned to applying the same methods to analyze biofluids such as serum, plasma and urine. For example, the number of proteins that have now been identified in blood rose exponentially to well over 1,000 in 2004. Though there are many several approaches, it is not yet clear as to which of these, or any other, method is the best one for biomarker discovery. It is likely that no single method will be universally used in the future to discover biomarkers, and the path chosen may greatly depend on the specific disease or previous knowledge of already acknowledged biomarkers.

Large-scale identification of bio-fluid proteomes:
Before comparative proteomics studies of biofluids for the purpose of identifying differentially abundant proteins that may serve as disease-specific biomarkers could proceed, methods had to be developed to allow for the comprehensive coverage of such proteomes. Plasma and serum represent the two most important biofluids in which the search for biomarkers has primarily been focused1 both of these samples, however, represent a challenge to proteomic characterization. On the positive side, they have a very high protein concentration (i.e. 60–80 mg/ml), so that only a very small aliquot would seemingly be needed for global mass spectrometry (MS)-based analysis. On the negative side, a great majority (i.e. 99 per cent) of this protein content comprises only 22 proteins. A reasonable hypothesis would suggest that disease specific biomarkers would be present within the 1 per cent of proteins that make up the low-abundance component of plasma or serum. Unfortunately, the dynamic range of protein concentration between the low- and high-abundance proteins is thought to span approximately ten orders of magnitude, which is much greater than the two orders of magnitude that MS
measurements are constrained to. Fortunately, investigators have explored many different fractionation methods prior to MS analysis to allow low abundance proteins to be identified in serum or plasma analysis. These fractionation methods have included immune depletion, chromatography and/ or electrophoresis, fittingly, because of its historically invaluable role in the development of proteomics, two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) fractionation played a major part in one of the first large-scale studies that showed the ability to identify hundreds of proteins within plasma[6] although 2D-PAGE has excellent resolution, the plasma sample still needed to be immune depleted of the highest-abundance proteins (i.e. albumin, haptoglobins, transferrins, transthyretin, \(\beta\)-1-antitrypsin, \(\beta\)-1-acid glycoprotein, hemopexin and \(\beta\)-2-macroglobulin) and sequentially fractionated by anion-exchange and size exclusion chromatography prior to its analysis by 2D-PAGE (Figure 1). This fractionation scheme resulted in a total of 74 2D-PAGE gels being run and the subsequent visualization of approximately 20,000 protein spots. Identification of many of these spots resulted in the net identification of 350 unique proteins, including interleukin-6, metallothionein II, cathepsins and peptide hormones, known to be present in serum at a concentration of less than 10 ng/ml.

Almost concurrently, while the above described immune depletion/ chromatography/2D-PAGE plasma analysis was being conducted, other investigators were conducting non-gel (or solution-based) methods to characterize plasma. One of the first studies utilized Ig depletion, multi-dimensional chromatography and MS to characterize the serum proteome[4] Serum was depleted of Igs and then digested with trypsin to create a complex mixture of peptides. These peptides were initially separated into 60 fractions using strong cation-exchange chromatography followed by reverse-phase liquid chromatography (RPLC)-MS/MS analysis of each fraction. Almost 500 proteins were identified using this method, including such species as prostate-specific antigen (PSA), which was believed to be present in the serum used in this study at a concentration of approximately 1 pg/ml. Many other global studies have subsequently been conducted with the aim of characterizing serum or plasma proteins, bringing the current state of the art coverage up to approximately 1,500 proteins. Developments in fractionation and MS instrumentation have been primarily responsible for this increase in the number of proteins that can be characterized within biofluids such as serum and plasma.

Investigation of sub-bio-fluid proteomes:
Although they comprise proteins and other biological molecules, biofluids are not typically thought of in the same context as cells or tissues. Cells and tissues are thought of as packages of biomolecules interacting in concert with one another to result in the accomplishment of a series of different processes. Biofluids, however, are typically considered to be ‘rivers’ of proteins that are transported through the body, acting independently, unless they encounter a cell or tissue to which they are supposed to signal for a physiological event to occur. The authors’ laboratory has recently taken a different perspective on how serum and plasma can be analyzed on a semi-global level for better characterization of low-abundance proteins, thereby increasing the chances of finding disease-specific biomarkers. In this study, it was hypothesized that proteins in the blood interact with one another, forming complexes of proteins through both specific and non-specific interactions. Obviously, this hypothesis is not too surprising, as albumin is known as a ‘sticky’ protein that acts as a transporter for blood components such as hormones, cytokines and lipoproteins[2] the second hypothesis was that by targeting and extracting high-abundance proteins, it would be possible to enrich for lower abundance proteins that complexed to these. Again, it is the authors’ belief that characterization of this low-abundance fraction optimizes the probability of finding novel disease biomarkers. To prove these hypotheses, a ‘sub-biofluid’ fractionation procedure was used, in which a series of antibodies directed against six high-abundance blood proteins (i.e. albumin, IgA, IgG, IgM, apolipoprotein and transferrin) were used specifically to capture these proteins, as well as proteins associated with each. After a non-stringent washing, the targeted protein, along with all of the species that bound to it, was removed from the antibody column. Since this mixture is ‘contaminated’ with single highly abundant protein (i.e. the antibody targeted protein), the sample was then diluted fivefold in buffered 20 per cent acetonitrile and then passed through a 30 kDa cutoff ultra filtration membrane to collect the low molecular weight species. These mixtures were then digested with trypsin and analyzed by RPLC-MS/MS (Figure 2). Over 200 unique proteins were found to be associated with these high-abundant proteins. Of these, 12 proteins that are currently utilized as experimental or clinical biomarkers were identified, including the known serum proteins PSA, pregnancy plasma protein A, meningioma-expressed antigen and dihydropyridine reductase (Table 1). Other clinically relevant biomarkers, such as bone morphogenetic protein 3b, prostate transglutaminase, paraneoplastic antigen MA1, glycosylasparaginase, coagulation factor VII precursor, ryanodine receptor 2 and acid sphingomyelinase-like phosphodiesterase 3a, that were identified in this study had not previously been
identified in other fully global analyses of blood proteome samples. Obviously, it is very difficult to reach any conclusions on the specificity of the interactions between these biomarkers and the high-abundance proteins in blood; however, this study was the first to show that a potential archive of diagnostic information may be bound to large, highly abundant circulatory proteins.

**Figure 1:** Proteome-wide characterization of human plasma using a combined immune depletion/chromatographic/two dimensional polyacrylamide gel electrophoresis (2D-PAGE) fractionation strategy followed by mass spectrometry identification. High-abundance protein-depleted serum was serially fractionated using anion and strong cation exchange chromatography, resulting in a total of 74 fractions that were separated and visualized on 2D-PAGE gels. Analysis of the accumulative 20,000 spots resulted in the identification of 350 unique proteins.

**Table 1:** Clinically relevant biomarkers that were to be associated with one or more highly abundant proteins in serum

<table>
<thead>
<tr>
<th>Protein</th>
<th>Association</th>
<th>Marker specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glycosylasparaginase</td>
<td>Albumin</td>
<td>I-cell disease</td>
</tr>
<tr>
<td>Paraneoplastic antigen MA1</td>
<td>IgA</td>
<td>Paraneoplastic neurological syndrome</td>
</tr>
<tr>
<td>Meningioma-expressed antigen 6/11</td>
<td>Albumin</td>
<td>Meningioma</td>
</tr>
<tr>
<td>Dihydropterdine reductase</td>
<td>Apolipoprotein</td>
<td>Tetrahydrobiopterin deficiency</td>
</tr>
<tr>
<td>Coagulation factor VII precursor</td>
<td>Albumin</td>
<td>Liver cirrhosis</td>
</tr>
<tr>
<td>Acid sphingomyelinase-like phosphodiesterase 3a</td>
<td>Apolipoprotein</td>
<td>Bladder tumours</td>
</tr>
<tr>
<td>Prostate transglutaminase</td>
<td>IgA, IgG</td>
<td>Prostate cancer</td>
</tr>
<tr>
<td>Pregnancy plasma protein A</td>
<td>IgG, IgM</td>
<td>Down’s syndrome</td>
</tr>
<tr>
<td>Hsc70-interacting protein</td>
<td>Albumin</td>
<td>Cell growth</td>
</tr>
<tr>
<td>Ryanodine receptor 2</td>
<td>Albumin</td>
<td>Diabetes</td>
</tr>
<tr>
<td>Bone morphogenetic protein 3b</td>
<td>IgA</td>
<td>Osteophytosis</td>
</tr>
<tr>
<td>Prostate-specific antigen</td>
<td>Albumin, IgG</td>
<td>Prostate cancer</td>
</tr>
</tbody>
</table>

Ig, immunoglobulin
Figure 2: Method for identification of low-abundance proteins bound to common, high-abundance proteins in serum. In this technique, an antibody is covalently coupled to protein G and is then used to extract a specific high-abundance protein (e.g., albumin, immunoglobulin G, transferrin etc) from serum. A low-stringency wash is used to enable the co-isolation of low-abundance proteins that are bound to the targeted protein. To remove the high-abundance protein, the mixture is passed through a 30 kDa ultra filtration membrane and then digested with trypsin. The resultant peptides are then identified by liquid chromatography coupled online with tandem mass spectrometry.

Narrowing down where to look in the haystack:
Discovery-driven biomarker identification, in which there is previous knowledge of what one is looking for, is analogous to looking for a needle in a haystack. Relying solely on fractionation and MS to identify disease-specific biomarkers is going to require studies in which samples from healthy and disease affected patients are quantitatively or qualitatively compared. Up to this point, MS-based study design has been focused on generating lists of proteins identified in control samples and comparing it with a list of proteins identified in samples obtained from disease-affected individuals. It is easy to envision the laboriousness of such approaches and the ‘data overload’ that can result. The process of discovering biomarkers would be expedited by finding ways to narrow down what part of the biofluid ‘haystack’ needs to be searched to locate the biomarker ‘needle’. Processes that could narrow down the section of the proteome that needs to be assayed to identify biomarkers include cell-based assays to localize a desired activity or sub cellular fractionation to isolate specific cellular organelles. While not all biomarkers may enable the use of orthogonal techniques, any test that can point to the fraction of a biofluid that contains a relevant biomarker is invaluable. This point is well illustrated in the identification of the anti-proliferative factor (APF), a urinary biomarker for patients with interstitial cystitis (IC). IC is a chronic bladder disorder that affects approximately one million Americans, yet there is no definitive diagnostic marker for IC. The functional presence of APF had previously been recognized, as the application of urine from IC patients to bladder epithelial cells inhibited their proliferation.
in vitro. By combining this cell-based assay with a series of chromatographic steps, a specific fraction that contained APF could be narrowed down and characterized by RPLC-MS/MS. As shown in panel (A) of Figure 3, the MS base peak chromatogram for the APF-active fraction is much simpler than that which would be seen for a non-fractionated urine sample. While four putative structures were determined through de novo sequencing of the tandem MS spectra corresponding to APF, each of these pointed towards APF as being a glycosylated nonapeptide. The correct sequence of APF was then determined by using a combination of BLAST searches and lectin-binding studies, thereby showing the biomarker to be a sialo glycopeptide in which the peptide domain had 100 per cent homology to a peptide contained within the sixth trans membrane domain of frizzled-8, a Wnt ligand receptor. Unfortunately, biomarker discovery is never complete at the discovery phase. Validation studies were then designed not only to show that the correct structure had been formulated but, more importantly, that the proposed molecule was present in samples obtained from IC patients but not from matched healthy controls.

![Figure 3](image-url)
Conclusions:
Science is in the infancy of a new revolution. While it can be argued that the human genome project was invaluable for the genomic information that resulted from it, perhaps as important is the new frame of mind that it has brought to biological science. In this new way of thinking, entire systems of genes, transcripts, proteins and metabolites are analyzed in single studies. The challenge is how to implement the most valuable information presented within these amounts of data. Biomarker discovery using MS-based proteomics is directly within the centre of this paradigm. While the discovery of APF as a biomarker for IC was able to rely on the use of a cell based assay, it is more likely that most biomarkers will not have the luxury of the use of such a test for their discovery. The most common study design relies on the immense data-gathering capabilities of MS as an aid to identifying proteomic differences between biofluids obtained from healthy and disease-affected patients. While there will be struggles ahead, the progress that has been made in the past two years, just in the characterization of biofluid proteomes, suggests that MS will be a leading technology in the discovery of disease-specific biomarkers in a very near future.

References: