REVIEW

Proteomic analysis of serum, plasma, and lymph for the identification of biomarkers

Zhaojing Meng and Timothy D. Veenstra

Laboratory of Proteomics and Analytical Technologies, SAIC-Frederick Inc., National Cancer Institute at Frederick, Frederick, MD, USA

Received: March 9, 2007
Revised: May 3, 2007
Accepted: May 9, 2007

Probably no topic has generated more excitement in the world of proteomics than the search for biomarkers. This excitement has been generated by two realities: the constant need for better biomarkers that can be used for disease diagnosis and prognosis, and the recent developments in proteomic technologies that are capable of scanning the individual proteins within varying complex clinical samples. Ideally a biomarker would be assayable from a noninvasively collected sample, therefore, much of the focus in proteomics has been on the analysis of biofluids such as serum, plasma, urine, cerebrospinal fluid, lymph, *etc.* While the discovery of biomarkers has been elusive, there have been many advances made in the understanding of the proteome content of various biofluids, and in the technologies used for their analysis, that continues to point the research community toward new methods for achieving the ultimate goal of identifying novel disease-specific biomarkers. In this review, we will describe and discuss many of the proteomic approaches taken in an attempt to find novel biomarkers in serum, plasma, and lymph.

Keywords:
Biomarker / Clinical diagnostics / Lymph / Plasma / Serum

1 Introduction

The human circulatory system is made up of the heart, blood, and over 100 000 kilometers of veins, arteries, and capillaries. No other biofluid has an intimacy with the body like blood has and therefore it is not surprising that it possesses such a richness of information concerning the overall pathophysiology of the patient. Unlike specific cell types, however, blood does not contain its own genome. Its genome can be considered as a compilation of the organism's genetic material, containing all of the variations (*i.e.*, mutations, single nucleotide polymorphisms, gene duplications, *etc.*) that are found in particular cells. Since it lacks a specific genome, it follows that blood does not have its own transcriptome. Rather it can potentially contain any portion of a transcript that is transcribed within any cell in the body. Likewise, the proteome of blood potentially contains portions of any of the proteins found within the organism's cell complement. A recent study comparing *N*-linked glycopeptides within cultured cells and solid tissues with plasma showed that numerous proteins from different cells and tissues are indeed present within this biofluid (Fig. 1) [1]. This study confirmed the prevailing hypothesis that blood contains proteins from a variety of different cells and tissues within the body and also substantiates the continued need for research into biofluid proteomics as a source of novel biomarkers.

The movement of substances to and from cells is critical for survival. In the human body, this transport function is carried out at a macroscopic level by the circulatory and lymphatic systems. The human circulatory system circulates approximately five liters of blood continuously throughout

Correspondence: Dr. Timothy D. Veenstra, SAIC-Frederick Inc., National Cancer Institute at Frederick, P.O. Box B, Frederick, MD 21702, USA
E-mail: veenstra@ncifcrf.gov
Fax: +1-301-846-6037

Abbreviations: ICAT, isotope-coded affinity tag; **LPS,** lipopolysaccharide; **SARS,** severe acute respiratory syndrome; **SCX,** strong cation exchange; **TBI,** traumatic brain injury

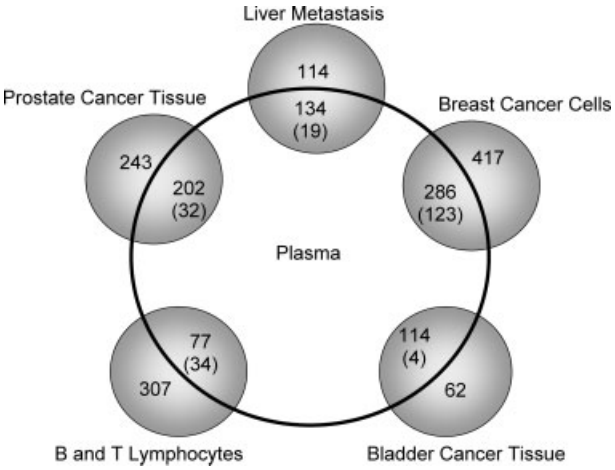

Figure 1. Identification of *N*-linked glycosites within various cells, solid tissues, and plasma. A combination of enrichment methods and MS was used to identify glycopeptides in plasma, B and T lymphocytes, a metastatic liver tumor, breast, bladder, and prostate cancer cells. A total of 1105 glycosites were identified in plasma alone. The unbracketed number inside the ring and within the gray circle represents the number of glycopeptides that were found within both plasma and the individual sample. The number outside the ring and within the gray circle represents the number of glycopeptides found within the individual sample but not within plasma. The number in parentheses represents the number of glycopeptides that were unique to the tumor tissue or cultured cell type [1].

the body of an average adult [2]. As blood enters capillaries, plasma leaks out to fill spaces between individual cells of tissue, becoming interstitial fluid. The delivery of nutrients to the cell is balanced by the transport of other components (*i.e.*, waste products, signaling hormones, *etc.*) back into the circulatory system. The balance between capillary oncotic pressure and hydrostatic pressure causes a slow increase in the volume of interstitial fluid. Approximately 90% of the fluid that enters the interstitial space enters back into the circulatory system by osmosis. The remaining 10% of the excess interstitial fluid diffuses into lymph capillaries and is returned back to circulation, after it has been processed within the lymph nodes, *via* the lymphatic system [3]. Once within the lymphatic system, the interstitial fluid is called lymph. Most lymph rejoins the circulatory system through the thoracic duct, the largest lymphatic vessel in the body.

Great effort has recently been exerted into characterizing the proteomes of various biofluids. The dominating reason is the hope that biomarkers indicative of a prevailing disease condition can be discovered. Much of the effort has focused on components of the circulatory system because no other biofluids possess such a broad range of characteristics that make it optimal for the discovery of protein biomarkers. First, as mentioned previously, no other fluid has complete intimacy with the body in the way blood has. In general, no cell in the human body is outside the diffusion distance of the circulatory system. Therefore, blood contains valuable information reflecting the specific physiological and pathological state of the whole human body. The extent of this information is reflected in the complexity of the proteomes of serum, and plasma. Second, blood is obtained through a relatively noninvasive procedure (*i.e.*, venipuncture). Third, the protein content of blood is very high. While other biofluids may better fulfill one of the above criteria (*e.g.*, urine is collected in an even less invasive manner), they do not completely satisfy the above characteristics in the manner that blood does. Because of the relatively noninvasive nature for its acquisition and the amount of information it can potentially provide, blood is routinely collected for biomarker screening and for monitoring the condition of the patient over long periods of time. Even though lymph does not encompass all of the same valuable characteristics, its close relationship with blood and the body's immune response makes it a potentially valuable source of novel disease-specific biomarkers.

2 How do serum, plasma, and lymph differ?

In most biomarker-driven proteomic studies, it is the plasma or serum portion of blood, rather than whole blood itself, that is analyzed. Plasma refers to the liquid component of blood and makes up about 45–55% of the total blood volume [4]. Blood cells, such as red and white blood cells (RBCs and WBCs) and platelets, are suspended within plasma. Plasma is collected by withdrawing blood in the presence of an anticoagulant (*e.g.*, heparin, sodium citrate, EDTA, *etc.*) and promptly centrifuged to remove the cellular elements. Serum is prepared by withdrawing blood in the absence of any anticoagulant, allowing for the formation of a fibrin clot [4]. Centrifugation is then used to remove blood cells and a large portion of the fibrinogen content *via* the fibrin clot.

The process of coagulation makes serum qualitatively different from plasma. At a macroscopic level, the protein concentration of serum is less than that of plasma; however, the differences have been shown to be on the order of only 3–4% [5]. This difference is largely a result of the removal of a large portion of the fibrinogen content of plasma in the form of the fibrin clot. Other proteins, however, are also removed by specific or nonspecific interactions within the fibrin clot. Conventional thinking would surmise that many coagulation factors are also removed in the preparation of serum. Actually factors IX, X, XI, and VII/VIIa are found within serum [6]. While its primary effect is the removal of the fibrin clot, coagulation involves platelet activation and coagulation cascades with many reactions occuring in the process. One study showed that the levels of platelet-secreted vascular endothelial growth factor (VEGF) are 230 ± 63 and 38 ± 8 pg/mL in serum and plasma of normal individuals, respectively [7]. In studies of patients suffering from thrombocytosis, in which their platelet count is substantially increased compared to matched healthy controls, VEGF

levels were also found to be much higher in serum than in plasma [8]. These results show that serum and plasma VEGF levels are affected by platelets, but more markedly so in the serum.

One of the most commonly asked questions is whether to use serum or plasma for biomarker discovery. The HUPO recommended based on its pilot phase of the Plasma Proteome Project (PPP) of 2002 the use of plasma over serum [9]. This recommendation was put forth because of the lower degree of *ex vivo* degradation observed in plasma samples that were analyzed using a variety of proteomic platforms. In addition, the recommended anticoagulants for plasma were citrate and EDTA, and not heparin. This recommendation is pretty obvious considering the molecular size and difficulty in removing heparin compared to citrate and EDTA. Heparin is highly charged and could interfere with the subsequent MS analysis, especially when a profiling method such as SELDI-TOF is being used. The choice of a sample type and preparation method has to be targeted to the specific biomarker discovery needs with a closely planned and controlled procedure. It has been reported that not only the sample choice (*e.g.*, serum or plasma) but also the sample-collection protocol (*e.g.*, type of collection tube) and the sample-processing procedure (*e.g.*, coagulation temperature, time allowed for coagulation, and anticoagulant used) could all bias the final results [10, 11]. Care must be taken when the archived samples are analyzed, as a recent HUPO study clearly showed how sample processing has a significant impact on the obtained results [9]. An important step in biomarker discovery will be the development of standardized methods that allow cross comparison of different studies.

While lymph is closely related to both plasma and serum it is not prepared from either of the two. As mentioned previously, lymph is made up of approximately 10% of the interstitial fluid that does not reenter circulation and is captured by the lymphatic system [3]. The lymphatic system is a major component of immune response system and is made up of a network of organs, lymph nodes, lymph ducts, and lymph vessels that transport lymph from tissues to the bloodstream. The tonsils, adenoids, spleen, and thymus are all considered a part of the lymphatic system. Lymph ranges from clear-to-white and contains RBCs, WBCs (primarily lymphocytes), as well as proteins and fats. Lymph is acquired from the patient by insertion of a cannula into the thoracic duct. The process of acquiring lymph is by no means routine and probably explains why it has not been the subject of as many proteomic investigations as serum and plasma.

3 The coupling of high-resolution MS and clinical science

The realization that MS-based technologies had the capability of identifying large numbers of proteins within complex proteomes was first shown through the combination of 2-DE and MS. During the mid 1990s, several laboratories combined the high-resolution separation capabilities of 2-DE with the high-throughput identification of MS to characterize a number of complex proteomes [12]. It was several years later that John Yates' laboratory showed the ability to circumvent 2-DE and use a combination of multidimensional fractionation and MS/MS analysis to identify almost 1500 proteins from *Saccharomyces cerevisiae* [13]. With the capability of identifying large numbers of proteins in a comparatively rapid manner realized, proteomics turned its efforts to the characterization of complex proteomes from a variety of different organisms and cell types.

While most of the early analytical focus in proteomics was on cultured cells and simple prokaryotic and eukaryotic organisms, in the early part of this decade a number of researchers including George Wright Jr., Daniel Chan, Sam Hanash and William Hancock amongst others, initiated clinical studies examining human biofluids. In 2002, a paper showing the ability to correctly diagnose serum samples obtained from women with ovarian cancer using simple TOF spectra obtained using a low-resolution mass spectrometer was published by Lance Liotta and Emanuel Petricoin [14]. While this study did not focus on broad-scale protein identification, it was limited to examining the low molecular weight fraction of the serum proteome, and the results remain extremely controversial, it created such a frenzy in the scientific community that many high-resolution MS/MS-based proteomic laboratories began focusing on methods for analyzing clinically important biofluids.

4 Approaches for targeting low-abundant proteins in serum/plasma

It was recognized early on, particularly in the analysis of serum and plasma, that the high dynamic range of protein concentrations found in these two fluids was going to be problematic for downstream MS analysis [15]. On the surface, serum and plasma seem to be the ideal clinical samples for MS-based proteomic analysis. They are relatively easy to obtain from the patient and have a very high protein concentration (*e.g.*, on the orders of tens of mg/mL). The protein concentration, however, is deceiving. Twenty-two proteins make up approximately 99% of the protein content of serum and plasma (Fig. 2). It is estimated that the protein concentrations in these samples span ten orders of magnitude and the prevailing thought is that specific disease biomarkers for diagnostic and prognostic purposes are most likely to remain within the very low concentration range. Considering that the dynamic range of a mass spectrometer is on the order of two orders of magnitude, it is easy to figure out that a straightforward LC-MS/MS analysis will result in the characterization of only the highest abundance, and probably least interesting, proteins. While strong cation exchange (SCX) prefractionation prior to a RP LC-MS/MS analysis has been shown to increase the ability to identify

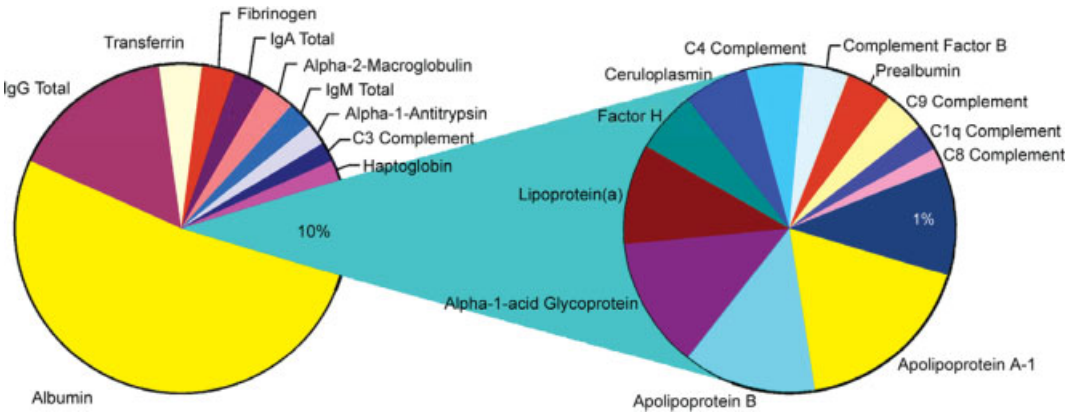

Figure 2. Graph showing the relative abundance of various proteins within serum. Twenty-two proteins make up roughly 99% of the protein concentration in serum.

low-abundant proteins in many proteomic studies [16, 17], this strategy alone is not sufficient to gain comprehensive coverage of the low-abundant proteins within biofluids.

It was quickly recognized that to effectively characterize serum or plasma was going to require methods to remove the high-abundant proteins prior to downstream analysis. One of the earliest approaches used to deplete high-abundant proteins was to pass a serum/plasma sample over Cibracon blue, a dye with a high affinity for albumin [18]. Albumin, as shown in Fig. 2, comprises approximately 50% of the protein content of serum/plasma. Recently, Agilent introduced the multiple affinity removal system (MARS) for the immunodepletion of six high-abundant proteins (*i.e.*, albumin, IgG, IgA, transferrin, haptoglobin, and alpha-1-antitrypsin) in serum/plasma [19]. Similar products have been developed, including a ProteoPrep 20 plasma immunodepletion kit from Sigma, and the Seppro™ MIXED12 IgY-based affinity LC column, for the depletion of the 12 highest abundant plasma proteins manufactured by GenWay Biotech [20]. The reproducibility and effectiveness of these products to deplete major proteins in serum/plasma samples have always been a concern. In fact, a recent study published the results of the reproducibility of a MARS column across serum samples from patients with prostate cancer. They found that the depletion of high-abundant proteins from all 250 serum samples was complete and reproducible, with a RSD below 7%, over a six week period [19]. A recent study comparing a series of sample preparation methods has also confirmed the effectiveness and robustness of immunoaffinity subtraction methods for simplifying the serum proteome prior to MS analysis [21]. Depletion of high-abundant proteins is now considered an essential sample-handling step in any serum/plasma study regardless of subsequent analytical strategies. There are always concerns, however, when using affinity based depletion strategies that potentially important biomarkers will be lost either through the possible "sponge" effect of the high-abundant proteins or by the nonspecific binding to the affinity column used. Indeed, studies have shown that proteins remain bound to the targeted high-abundant proteins during their depletion [22, 23]. Moreover, a major protein depletion alone certainly was not enough to deal with the dynamic range problem.

Besides the affinity depletion approaches, alternative approaches have been applied to target and isolate a subproteome of the serum/plasma in order to reduce the sample complexity and improve low-abundant protein characterization. One such approach has utilized hydrazide chemistry to capture and enrich glycoproteins onto a solid support and eventually release *N*-linked glycosylated peptides using *N*-glycosidase [1, 24]. Glycosylation plays a significant role in modulating the function and physiology of body and aberrant glycosylation has been implicated in many diseases. Since most secreted proteins are glycosylated, enriching for this class of peptides not only reduces serum/plasma sample complexity but also provides a targeted approach for biomarker discovery. This glycopeptide-targeted approach is capable of identifying hundreds of glycopeptides in a single analysis. Another approach is to apply a reversible capture release cysteinyl-peptide enrichment method using thiopropyl-sepharose 6B thiol affinity resin to reduce serum/plasma sample complexity [25]. This method is most effective, however, when used in combination with albumin depletion as this protein contains a large number of cysteinyl residues. This technology has shown the capability of identifying and quantitating over 600 proteins in a single LC-MS/MS run.

5 Identification of the serum, plasma, and lymph proteomes

To achieve the dynamic range measurements needed for serum/plasma samples, it has become a common practice to use a combination of strategies of depletion and fractionation strategies. Fittingly, one of the first large scale studies that showed the ability to identify hundreds of proteins within a biofluid, in this case serum, incorporated 2-DE with

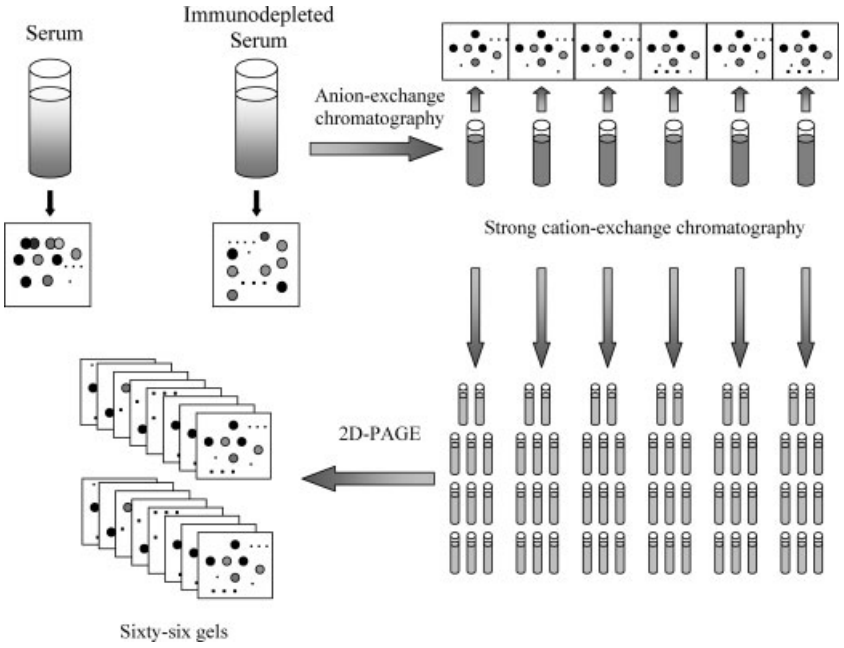

Figure 3. Characterization of the human serum proteome using immunodepletion/chromatographic/2-DE fractionation strategy followed by MS identification. Serum, in which the high-abundant proteins had been immunodepleted, was fractionated using anion-exchange and size-exclusion chromatography resulting in a total of 74 fractions that were separated and visualized on 2-DE gels. Raw and immunodepleted serum were directly separated on two other gels. Analysis of the accumulative 20 000 spots resulted in the identification of 350 unique proteins [26].

MS identification. (Fig. 3) [26]. Recognizing its large dynamic range of protein concentration, the serum sample was immunodepleted to remove the most abundant proteins (*i.e.*, albumin, IgG, haptoglobins, transferrins, transthyretin, α-1-antitrypsin, α-1-acid glycoprotein, hemopexin, and α-2-macroglobulin). The remaining proteins were separated into 74 fractions using sequential anion-exchange and size-exclusion chromatography. Each of these fractions was run individually on a 2-DE gel. Coomassie staining of the gels resulted in approximately 20 000 individual protein spots. Removal of redundant spots by the analysis of the visual images still left 3700 unique spots. Analysis of these spots by MALDI-TOF and/or LC-MS/MS resulted in the identification of 1800 of these spots, which could be correlated to 325 unique proteins. So what did they find in serum? Almost 39% of the proteins identified were known to be localized within the circulatory system, while 35% represented intracellular proteins that are hypothesized to leak into circulation. Proteins that are known to reside on the cell surface made up just over 6% of the total number of unique protein identifications. Not surprisingly, considering the amount of fractionation that was conducted, several proteins with known serum concentrations less than 10 ng/mL (*e.g.*, interleukin-6, metallothionein II, cathepsins, and various peptide hormones) were identified.

Almost concurrent with the above study, the laboratories of Richard D. Smith and Joel Pounds at Pacific Northwest National Laboratories were investigating the human serum proteome using multidimensional fractionation of a serum tryptic digestate combined with MS/MS identification (Fig. 4) [27]. As done with the above 2-DE-based study, serum was immunodepleted, however, only for Igs and not several of the other major high-abundant proteins. This immunodepleted sample was fractionated into 60 aliquots using a SCX LC. Each of these aliquots was analyzed using a microcapillary RP LC coupled on-line with tandem MS. This solution-based (or "shotgun") method resulted in the identification of 490 unique proteins (*cf.* 325 in the 2-DE analysis). As with the 2-DE study described above, many of the expected circulatory proteins were identified as well as those originating from cells and tissues throughout the body. Several very low-abundant proteins, such as prostate-specific antigen (PSA), which are believed to be present at concentration in ng/mL range in the serum sample, were identified using this method. Both of these studies illustrated that the current technology is sensitive enough to detect low-abundant proteins in serum and plasma. Further developments in technology to deal with the serum and plasma sample dynamic range without losing the low-abundant proteins or a more targeted sample analysis are required for biomarker discovery.

A comparison between these two serum analyses presents some obvious advantages/disadvantages for either strategy. The 2-DE strategy appears to be extremely laborious, requiring significant prefractionation prior to running 74 gels. This fractionation is followed by the selection of 3700 protein spots that are required in-gel tryptic digestion and MS analysis by MALDI-TOF, with additional LC-MS/MS analysis in cases where peptide mapping was unsuccessful. The 2-DE method, however, does provide an inherent protein quantitation if comparative studies are conducted, through the staining of the proteins fractionated within the gel. In addition, isoforms originating from differential PTMs such as glycosylation can be observed for individual proteins. The multidimensional LC approach is less laborious, requiring a single tryptic digest and tens of LC-MS/MS analyses to

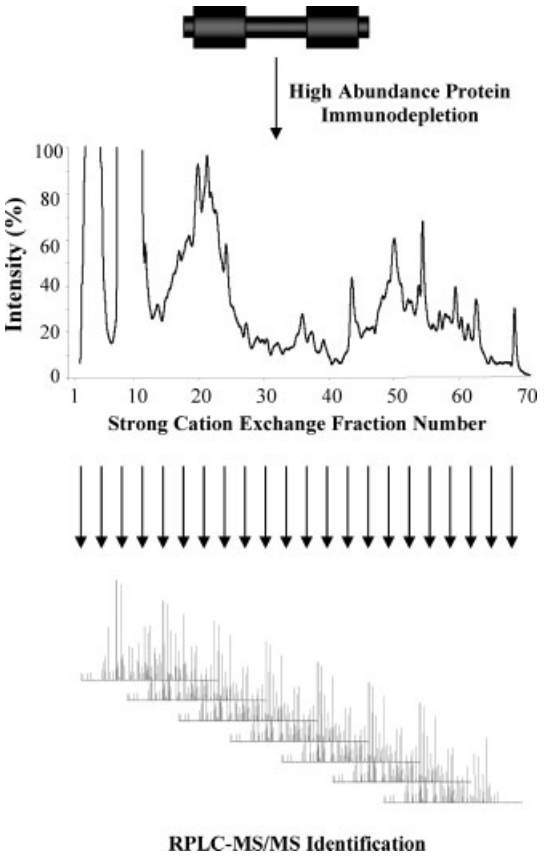

Figure 4. Schematic showing standard nongel method to analyze complex biofluids in which immunodepleted serum or plasma is initially separated using a SCX chromatography. Each of these fractions is subsequently analyzed using RP LC coupled directly on-line with MS/MS.

identify the proteins present within the serum. This method appears to be more sensitive than the 2-DE method for the detection of low-abundant proteins. The solution-based fractionation method does not, however, provide the individual protein coverage capable through the analysis of individual proteins fractionated by 2-DE. In addition, the solution-based fractionation strategy lacks direct quantitative comparison capabilities, however, algorithms such as spectral counting or an exponentially modified protein abundance index (emPAI) [28] have been employed recently to quantify protein data acquired using such a method.

After the publication of these two studies, many other laboratories began their own exploration of the serum proteome. A recent study in human plasma proteome HUPO pilot phase project applied a Hi-D separation strategy using major protein depletion, IEF, one-dimensional gel electrophoresis (1D-SDS) to fractionate the proteins before they were digested and analyzed by RPLC-MS/MS on a linear IT mass spectrometer (LTQ) [29]. A total of 159 samples were analyzed by RPLC-MS/MS. A detection dynamic range of nine orders of magnitude with 575 and 2890 proteins identified from plasma and serum, respectively, were reported. Of these identified proteins, 16 are known to be present in the pg/mL to ng/mL range. A number of subsequent studies utilized the SCX-LC with RPLC-MS/MS approach for both plasma and serum and the technology rapidly matures to a point that it was not uncommon to identify thousands of proteins within these biofluids. Obviously there will be differences in the number of proteins identified in studies, which is generally an effect of the sample processing methods used, but can also be dependent on the criteria and stringency used for MS/MS identification. When using SEQUEST to analyze raw MSW/MS data, a number of different parameters such as enzyme constraint, cross-correlation (X_{corr}), and delta correlation (ΔC_n) score have a significant impact on the number of peptides that are considered as "confident" identifications. An excellent example is presented in a study conducted in Dr. Richard Smith's laboratory [30]. The same dataset of MS/MS spectra obtained from human plasma was analyzed using a variety of different enzyme constraints and X_{corr} and ΔC_n values used in other publications. Their results showed that the number of peptide and protein identifications ranged from 2912 to 3935 and 880 to 1682, respectively, depending on the criteria chosen.

Recently, an extensive reference plasma proteome database from trauma patient has been established using the combination of major protein depletion, target protein enrichment, and multidimensional LC [31]. The crude plasma was processed and analyzed as illustrated in Fig. 5. After removal of 12 high-abundant proteins, the sample was split into two aliquots. One of the aliquots was digested with trypsin, and a thiol affinity resin was added to this mixture allowing for the enrichment of cysteinyl-containing peptides. The other aliquot was oxidized by period and the glycoproteins were covalently coupled to hydrazide beads. These proteins were then digested with trypsin and the released peptides were isolated. The *N*-glycopeptides that remained bound to the beads were released using PNGase F. All four of the fractions (*i.e.*, noncysteinyl peptides, cysteinyl-containing peptides, nonglycopeptides, and *N*-glycopeptides) were then individually separated using SCX into 30 fractions. Each fraction was analyzed using RPLC-MS/MS. Gene Ontology (GO) analysis revealed the identification of a large number of inflammation and immune response-related proteins in this sample. There were a total of 22 267 unique peptides identified in this extensive study corresponding to 3654 nonredundant proteins. The various fractionation strategies afforded the identification of proteins over a dynamic range of protein concentration greater than seven orders of magnitude. Many low-abundant proteins including 78 cytokines and cytokine receptors (such as tumor necrosis factor receptor, interleukin, vascular endothelial growth factor, and transforming growth factor-β, *etc.*) as well as 136 cell differentiation molecules were also identified using this method. While this method is obviously laborious and quite sophisticated, it provides an effective

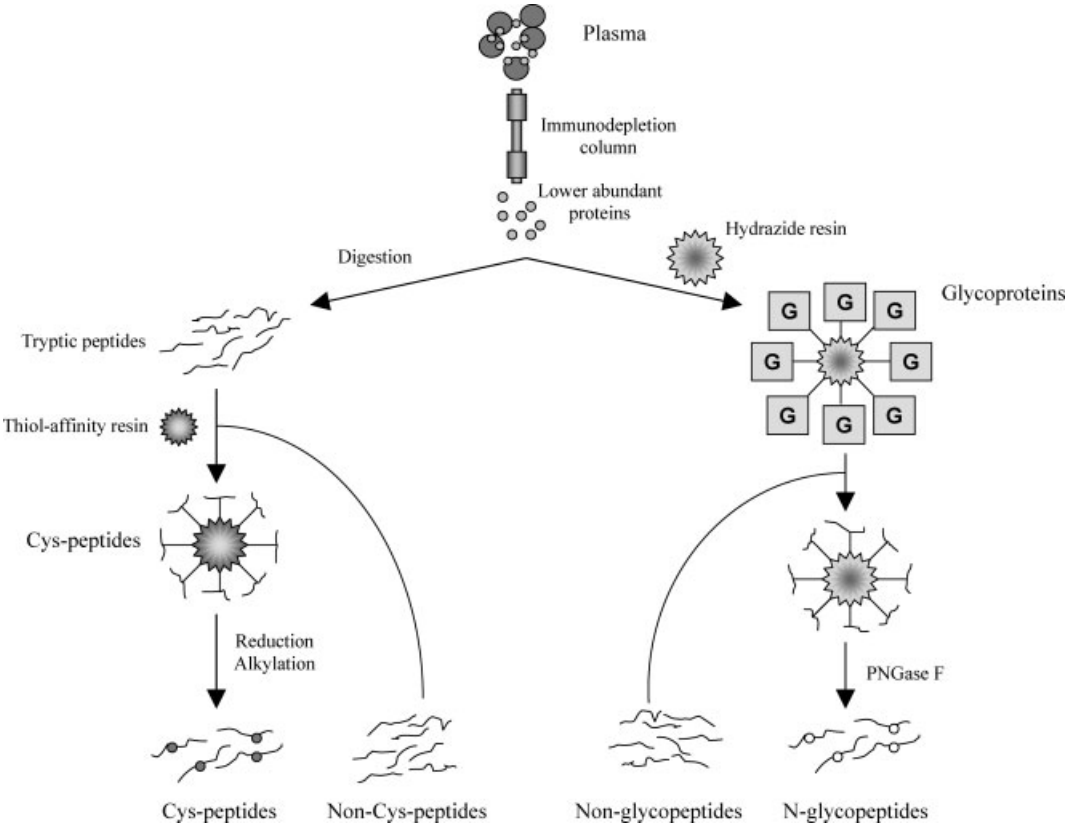

Figure 5. Schematic representation of the sample prefractionation processes used to characterize the plasma proteome of plasma obtained from patients with TBI [31].

illustration into the importance of being able to observe low-abundant proteins if we anticipate using proteomic technology to identify novel biomarkers.

Although most multidimensional LC methods that characterize biofluids are based on the separation of peptides, a recent study fractionated intact proteins according to their p*I*s (8.5 to 4.0) followed by their hydrophobicity. This sequential chromato-focusing/RP chromatography system is commercially available from Beckman Coulter, under the brand name ProteomeLab PF2D system (http://www.beckmancoulter.com/products/instrument/protein/proteomelab_pf2d_dcr.asp). The intact protein fractions are then digested using trypsin and analyzed by LC-ESI-MS/MS or MALDI-MS for protein identification. Application of this method using an albumin- and Ig-depleted serum sample from a healthy individual resulted in the identification of 150 proteins [32]. Included in these identifications were proteins spanning three orders of magnitude in concentration (*e.g.*, coagulation factor XIII and troponin T, which are present at μg/mL and ng/mL quantities, respectively). This strategy of fractionating intact proteins offers a greater opportunity of identifying various protein isoforms compared to methods that predigest the proteome samples into peptides prior to separation.

While most of the developments in finding more effective ways to characterize the proteomes of complex biofluids have focused on sample preparation, novel MS instrumentation methods are also being pursued. A recent application using ion mobility-MS (IMS-MS) in combination with multidimensional LC to characterize the plasma proteome was able to yield 731 highly confident peptide identifications in merely 3.3 h without the need for high-abundant protein depletion [33]. Even though there are limitations to this emerging technology, IMS-MS undoubtedly shows a great potential as a new MS-based approach for high-throughput serum/plasma proteome analysis in the future.

Lymph has been analyzed in only a few reports. In one of the first reports, normal ovine lymph was compared to plasma [34]. Proteins from both samples were analyzed using 2-DE. Both the lymph and plasma gels were dominated by albumin. Other plasma proteins that were observed in lymph included fibrinogen α- and β-chains, IgG heavy chain, serotransferrin precursor, lactoferrin, and apolipoprotein A-1. Two proteins, glial fibrillary astrocyte acidic protein and neutrophil cytosol factor-1, were found to be differentially abundant in lymph, showing that this biofluid is simply more than just an ultrafiltrate of plasma. Even though the process of acquiring lymph is not trivial, it contains a lower

concentration of large proteins since bigger plasma proteins, including albumin, do not readily pass through capillary walls into the interstitial fluid. This feature along with its close relationship to the immune response makes lymph a very interesting and informative proteome. The popular strategies that have been applied to serum/plasma sample studies described above could certainly help to better define proteomic differences between lymph and plasma/serum.

Unfortunately these studies did not provide the "biomarker goldmine" that was anticipated; however, they did reveal the complexity that the proteomic community was facing as it moved forward in the search for diagnostic and therapeutic biomarkers. While the proteins identified in these many studies could be grouped using a number of classification categories (*e.g.*, localization, molecular function, *etc.*), one thing was obvious, serum and plasma not only contained the expected circulatory proteins, but also proteins from every conceivable source (*i.e.*, cell surface, cell nucleus, cell cytoplasm, mitochondria, *etc.*) in the body.

6 Comparative analysis strategies for biomarker detection

There has probably been no more active field in proteomics over the past few years than the search for biomarkers. A simple search of PubMed using the terms "serum" and "proteomics" gives 673 citations since the year 2000, with approximately one-third of these being published since the beginning of 2006. Within these citations are studies that have a variety of different aims and use a number of different technologies. A detailed description of every different technology would fill this entire journal edition by itself. The relative importance of each study highlighted below is left up to the individual reader; however, we have endeavored to select examples that illustrate the breadth of techniques used to identify biomarkers within serum, plasma, or lymph.

Obviously, the comprehensive identification of proteins in a single clinical sample is not going to reveal useful biomarkers. Such studies require some type of comparisons to be made between samples obtained from different populations (*e.g.*, healthy *versus* disease-affected patients). As with any proteomic comparison, 2-DE remains a stalwart in quantitative comparison of biofluids. A 2-DE approach was recently applied to compare plasma samples obtained from patients with severe acute respiratory syndrome (SARS) and healthy individuals [35]. Twenty-two plasma samples from four different SARS patients were separated by 2-DE using a narrow range IPG strip (pH 4–7) and the resulting profiles compared to those obtained from six healthy plasma samples. Seven proteins were exclusively present in the 22 SARS samples. Eight additional spots were up-regulated in all 22 SARS patients compared to the healthy controls. Many of the proteins up-regulated in plasma from SARS patients can be classified as acute phase proteins (APP) that are produced as a consequence of serial cascades initiated by the SARS-coronavirus infection. Interestingly, the intracellular, antioxidant protein peroxiredoxin II was found to be up-regulated in all of the 22 SARS plasma samples. In a separate validation study, peroxiredoxin II was found in the plasma of approximately 36% of SARS patients, but only 10% of patients with fever. This rate of detection is higher than that found in human immunodeficiency virus (HIV) patients, suggesting that peroxiredoxin II may function as a useful serum biomarker for SARS infection.

In addition to changes in protein abundance, differences in specific PTMs are also critical in disease pathology. A recent study evaluated total protein, glycoprotein, and phosphorylated protein difference between ovarian cancer (OVC) patients and healthy controls [36]. Plasma samples from five OVC patients and five healthy controls were used for the study. Each pooled plasma sample was first depleted of the top six proteins and then analyzed by 2-DE in triplicates to be stained with stains specific for total protein, phosphorylation, and glycosylation respectively. A phosphorylated isoform of fibrinogen-α-chain was found up-regulated in this study, which agrees with an early low-molecular weight serum study of OVC patients from the same group.

Nongel based approaches for conducting comparative proteomic analyses of samples such as cell and tissue lysates have been widely applied, however, many of these methods that require stable-isotope labeling are not inherently useful for a comparative analysis of human biofluid samples. For instance, metabolic labeling with heavy isotopes while not impossible, as shown by the creation of a heavy-isotope labeled rat [37], may not be practical for this type of analysis in which many samples from different subjects need to be compared. In addition, isotope-coded affinity tag (ICAT)-labeling, which has been used in numerous studies comparing proteomes of cell cultures and tissues, has been used on a limited basis in the comparison of biofluids, except for CSF. A recent study, however, used ICAT-labeling to measure changes in protein abundance observed in serum obtained from pediatric patients with severe traumatic brain injury (TBI) [38]. Samples from six patients (heavy ICAT-labeled) were compared to a pooled sample of healthy adults (light ICAT-labeled). All samples were depleted of albumin and IgG prior to ICAT-labeling. A total of 95 proteins were found to be differentially abundant in the TBI serum samples compared to the pooled control. Most of the identified differentially expressed proteins are known to be involved in inflammation, innate immunity, and early stress/defense response. These proteins included several low-abundant proteins such as Toll receptors, signaling kinases, serine/threonine- protein kinases, transcription factors (serum response factor, golgin 45, myocyte-specific enhancer factor 2B), proteases (pappalysin-2 precursor, MMP-9), and proteins involved in response to oxidative-stress. The global changes in serum protein expression in TBI patients indicated a massive defense response with the most prominent response being

the recruitment of proteins involved in inflammatory and immune pathways. Several proteins that can potentially be localized to the brain were quantitatively measured in this study. Many of these, such as γ-enolase, amyloid β4 precursor, α-spectrin, and cleaved microtubule-associated protein tau, which have been previously detected in serum or CSF from TBI, or other types of brain injury, were found at increased levels in pediatric TBI patients. This study showed that ICAT-labeling can be useful in the comparative analysis of biofluid samples.

While there have also been studies utilizing $^{18}O/^{16}O$ trypsin-mediated isotopic labeling, comparative proteomics of biofluids have typically been limited to spectral counting studies where the number of identified peptides, or a peptide's peak area, is used as a measure of a protein's relative abundance compared to another sample. An excellent example of these methods was the study published by Richard Smith's lab in which peptide peak areas and the number of peptide identifications from 2D-LC-MS/MS analyses were used to garner a quantitative comparison of protein abundances between plasma samples obtained from a human subject prior to (untreated) and 9 h after lipopolysaccharide (LPS) administration (treated) [39]. LPS is an endotoxin released by Gram-negative bacteria that is known to induce inflammatory reactions, such as cytokine production, cell migration, and production of acute-phase proteins. This study sought to quantitate changes in the acute phase plasma proteome in response to the LPS administration. The untreated and LPS-treated plasma samples were digested with trypsin and each sample was fractionated using a SCX chromatography. A total of 50 fractions were collected for each sample and each of these was analyzed by RPLC-MS/MS. Some of the SCX fractions that had a high peptide content were run twice, resulting in a total of 148 RPLC-MS/MS analyses. Combining both analyses (*i.e.*, treated and nontreated) resulted in a total of 804 unique plasma proteins (not including IgGs) being identified from 5176 unique peptides. Of these, 83% (669 proteins) were identified by at least two unique tryptic peptides.

To determine if the number of peptide identifications for each protein could be used in a quantitative manner, the group plotted the number of peptides identified for 74 specific proteins against their literature-documented concentration in plasma (Fig. 6) [39]. In general, the correlation was quite good suggesting peptide hit number is at least semi-quantitative. The group also compared the peak areas for peptides that were identified in both samples and used this ratio, along with the number of peptide hits, to identify proteins that were differentially abundant in LPS-treated plasma. A number of proteins were found to be significantly increased in concentration following LPS administration. Amongst these included several inflammatory or acute-phase response proteins such as LPS-binding protein, LPS-responsive and beige-like anchor protein, C-reactive protein, serum amyloid A and A2, hepatocyte growth factor activator, and von Willebrand factor. As shown in Table 1, eight out of the nine proteins listed for which a protein abundance ratio was determined showed an increase in concentration following LPS administration by both the protein abundance ratios and the ratios of peptide hits. The two computational approaches, however, are generally complementary as many of the up-regulated proteins were identified in only one of the two methods. This study was one of the first to show that signal intensity and peptide hit count could be used to quantitatively compare protein abundances in biofluids analyzed by LC-MS/MS. Presently, most nongel based comparative studies of serum and plasma are conducted using either of these two computational approaches to measure the relative quantitation of proteins in two or more samples.

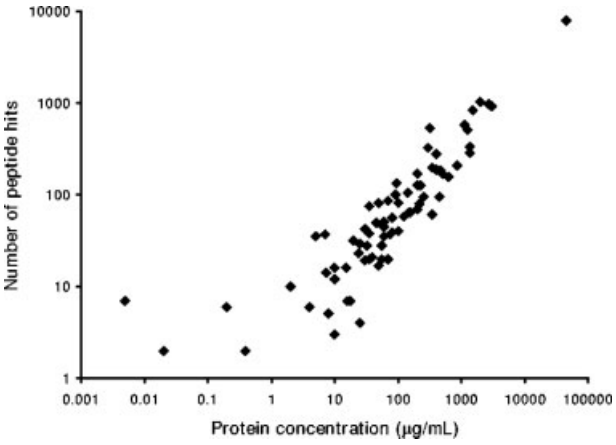

Figure 6. Correlation between the number of peptides identified for specific proteins during a multidimensional fractionation/MS/MS analysis of plasma compared to their documented concentration [39].

Table 1. Comparison of the ratio of peptide hits and relative abundance ratio (determined by measuring peak areas) for nine proteins observed to be up-regulated in the comparison of plasma taken from a patient prior to (untreated) and after treatment with LPS [39]

Protein	Ratio of peptide hits (treated/untreated)	Abundance ratio (treated/untreated)
Serum amyloid A	2.3	5.9
Serum amyloid A2	4	4.3
Hepatocyte growth factor activator	4	3.8
LPS binding protein	3.9	2.6
von Willebrand factor	4.3	1.1
KIAA1009 protein	7	2.9
Leucine-rich a-2 glycoprotein	1.5	2.87
KIAA1301 protein	4	2.87
NADH oxidase	2.6	1.95

7 Conclusions

A simple search of the literature will reveal the enormity of resources that have been spent to search for biomarkers, particularly in serum and plasma. If success is gauged by the number of *validated* biomarkers identified using MS-based methods, the result looks pretty bleak. However, if we examine the steps in the identification of clinically useful biomarkers, the impact of proteomic developments is readily obvious. The identification of novel biomarkers begins with discovery. The purpose of the discovery phase is to identify proteins that possess some characteristic (*e.g.*, abundance difference) that is different between samples obtained from disease-afflicted patients and healthy controls. This phase is usually conducted through the comparison of only a few (*e.g.*, 10–100) clinical samples. A useful discovery study will provide many potentially useful biomarkers. In general, it will not be possible to move all of the potential biomarker candidates forward into the validation phase, where it may be necessary to examine thousands of samples. Therefore, key decisions need to be made to determine which biomarker candidates identified in the discovery phase will be interrogated in the validation phase. Those proteins that survive validation then become targets for assay development to specifically measure their presence in clinical samples. Proteomic studies, as described above, have not had a major impact in the validation phase. Their real impact at the discovery phase, however, is without question. Biofluid proteomics as it is practiced today is a science in its infancy, yet the rate at which it has grown is astounding. It was not more than 5 years ago that only a handful of laboratories were capable of identifying a few hundred proteins in serum or plasma. Today the ability to identify thousands of proteins, as well as hundreds of differences between disease-afflicted and control samples are almost commonplace.

There are two major challenges facing clinical proteomics as it attempts to discover novel biomarkers. A simple reading of the literature, including this review, shows that many of the approaches that utilize MS/MS to characterize biofluids lack the throughput necessary to survey the number of clinical samples required to make any type of solid conclusion about the potential of any protein becoming a useful biomarker. The challenge will be to develop high throughput MS platforms that enable direct protein identification while having the capability of analyzing hundreds of clinical samples in a reasonable time frame. Fortunately, the speed at which current mass spectrometers are capable of performing MS/MS experiments makes this challenge attainable. The next challenge in proteomics will be to refine the decision point to increase the success rate at which candidates are validated as useful biomarkers. To produce a clinically useful biomarker requires four phases; discovery, qualification, verification, and validation with assay development [40]. These stages require the analysis of increasing number of samples, but fewer analytes within each sample. MS fits well within the discovery phase as it is able to measure hundreds of differences between clinical samples. Although the discovery phase only requires on the order of tens of samples to be analyzed, this number is still time consuming with current MS-based proteomic studies. As the throughput of MS instrumentation continues to increase, the ability to survey a greater number of samples, and repeatedly observe specific differences in protein abundances, will increase the confidence in selecting which differentially abundant proteins have the greatest chance of eventually becoming validated for clinical use.

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under Contract NO1-CO-12400. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organization imply endorsement by the United States Government.

8 References

[1] Zhang, H., Liu, A. Y., Loriaux, P., Wollshceid, B. *et al.*, *Mol. Cell. Proteomics* 2007, *6*, 64–71.

[2] Cameron, J. R., Skofronick, J. G., Roderick, M., Grant, R. M., *Physics of the Body*. 2nd Edn., Medical Physics Publishing, Madison, WI 1999, p. 182.

[3] Olszewski, W. L., *Lymphat. Res. Biol.* 2003, *1*, 11–21.

[4] Luque-Garcia, J. L., Neubert, T. A., *J. Chromatogr. A* 2007, *1153*, 259–276.

[5] Lum, G., Gambino, S. R., *Am. J. Clin. Pathol.* 1974, *61*, 108–113.

[6] Kimball, D. B., Rickles, F. R., Gockerman, J. P., Hattler, B. G. *et al.*, *J. Lab. Clin. Med.* 1976, *87*, 868–881.

[7] Spence, G. M., Graham, A. N., Mulholland, K. *et al.*, *Int. J. Biol. Markers* 2002, *17*, 119–124.

[8] Benoy, I., Salgado, R., Colpaert, C., Weytjens, R. *et al.*, *Clin. Breast Cancer* 2002, *2*, 311–315.

[9] Rai, A. J., Gelfand, C. A., Haywood, B. C., Warunek, D. J. *et al.*, *Proteomics* 2005, *5*, 3262–3277.

[10] Teahan, O., Gamble, S., Holmes, E., Waxmann, J. *et al.*, *Anal. Chem.* 2006, *78*, 4307–4318.

[11] Rai, A. J., Vitzthum, F., *Expert Rev. Proteomics* 2006, *3*, 409–426.

[12] Nordhoff, E., Egelhofer, V., Giavalisco, P., Eickhoff, H. *et al.*, *Electrophoresis* 2001, *22*, 2844–2855.

[13] Washburn, M. P., Wolters, D., Yates, J. R., III, *Nat. Biotechnol.* 2001, *19*, 242–247.

[14] Petricoin, E. F., Ardekani, A. M., Hitt, B. A., Levine, P. J. *et al.*, *Lancet.* 2002, *359*, 572–577.

[15] Anderson, L., *J. Physiol.* 2005, *563*, 23–60.

[16] Fujii, K., Nakano, T., Kawamura, T., Usui, F. *et al.*, *J. Proteome Res.* 2004, *3*, 712–718.

[17] Jin, W. H., Dai, J., Li, S. J., Xia, Q. C., *J. Proteome Res.* 2005, *4*, 613–619.

[18] Zolotarjova, N., Martosella, J., Nicol, G., Bailey, J. *et al.*, *Proteomics* 2005, *5*, 3304–3313.

[19] Darde, V. M., Barderas, M. G., Vivanco, F., *Methods Mol. Biol.* 2007, *357*, 351–364.

[20] Gong, Y., Li, X., Yang, B., Ying, W. *et al.*, *J. Proteome Res.* 2006, *5*, 1379–1387.

[21] Whiteaker, J. R., Zhang, H., Eng, J. K., Fang, R. *et al.*, *J. Proteome Res.* 2007, *6*, 828–836.

[22] Gundry, R. L., Fu, Q., Jelinek, C. A., Van Eyk, J. E., Cotter, R. J., *Proteomics Clin. Appl.* 2007, *1*, 73–88.

[23] Zhou, M., Lucas, D. A., Chan, K. C., Issaq, H. J. *et al.*, *Electrophoresis* 2004, *25*, 1289–1298.

[24] Zhang, H., Li, X.-J., Martin, D. B., Aebersold, R., *Nat. Biotechnol.* 2003, *21*, 660–666.

[25] Liu, T., Qian, W. J., Strittmatter, E. F., Camp, D. G. *et al.*, *Anal. Chem.* 2004, *76*, 5345–5353.

[26] Pieper, R., Gatlin, C. L., Makusky, A. J., Russo, P. S. *et al.*, *Proteomics* 2003, *3*, 1345–1364.

[27] Adkins, J. N., Varnum, S. M., Auberry, K. J., Moore, R. J. *et al.*, *Mol. Cell. Proteomics* 2002, *1*, 947–955.

[28] Ishihama, Y., Oda, Y., Tabata, T., Sato, T. *et al.*, *Mol. Cell. Proteomics* 2005, *4*, 1265–1272.

[29] Tang, H. Y., Ali-Khan, N., Echan, L. A., Levenkova, N. *et al.*, *Proteomics* 2005, *5*, 3329–3342.

[30] Shen, Y., Jacobs, J. M., Camp, D. G., Fan, R. *et al.*, *Anal. Chem.* 2004, *76*, 1134–1144.

[31] Liu, T., Qian, W.-J.,Gritsenko, M. A., Xiao, W. *et al.*, *Mol. Cell. Proteomics* 2006, *5*, 1899–1913.

[32] Sheng, S., Chen, D., Van Eyk, J. E., *Mol. Cell. Proteomics* 2006, *5*, 26–34.

[33] Valentine, S. J., Plascencia, M. D., Liu, X., Krishnan, M. *et al.*, *J. Proteome Res.* 2006, *5*, 2977–2984.

[34] Leak, L. V., Liotta, L., Krutzsch, H., Jones, M. *et al.*, *Proteomics* 2004, *4*, 753–765.

[35] Chen, J. H., Chang, Y. W., Yao, C. W., Chiueh, T. S. *et al.*, *Proc. Natl. Acad. Sci. USA* 2004, *101*, 17039–17044.

[36] Ogata, Y., Hepplmann, C. J., Charlesworth, M. C., Madden, B. J. *et al.*, *J. Proteome Res.* 2006, *5*, 3318–3325.

[37] Wu, C. C., MacCoss, M. J., Howell, K. E., Matthews, D. E. *et al.*, *Anal. Chem.* 2004, *76*, 4951–4959.

[38] Haqqani, A. S., Hutchison, J. S., Ward, R., Stanimirovic, D. B., *J. Neurotrauma* 2007, *24*, 54–74.

[39] Qian, W. J., Jacobs, J. M., Camp, D. G., II, Monroe, M. E. *et al.*, *Proteomics* 2005, *5*, 572–584.

[40] Rifai, N., Gillette, M. A., Carr, S. A., *Nat. Biotechnol.* 2006, *24*, 971–983.

