Available online at www.sciencedirect.com
ScienceDirect
Journal of Chromatography B, 847 (2007) 3–11
JOURNAL OF CHROMATOGRAPHY B
www.elsevier.com/locate/chromb

Review

Global and targeted quantitative proteomics for biomarker discovery☆

Timothy D. Veenstra*

SAIC-Frederick Inc., National Cancer Institute at Frederick, P.O. Box B, Frederick, MD 21702, United States

Received 19 June 2006; accepted 3 September 2006
Available online 4 October 2006

Abstract

The extraordinary developments made in proteomic technologies in the past decade have enabled investigators to consider designing studies to search for diagnostic and therapeutic biomarkers by scanning complex proteome samples using unbiased methods. The major technology driving these studies is mass spectrometry (MS). The basic premises of most biomarker discovery studies is to use the high data-gathering capabilities of MS to compare biological samples obtained from healthy and disease-afflicted patients and identify proteins that are differentially abundant between the two specimen. To meet the need to compare the abundance of proteins in different samples, a number of quantitative approaches have been developed. In this article, many of these will be described with an emphasis on their advantageous and disadvantageous for the discovery of clinically useful biomarkers.

Keywords: Biomarker discovery; Mass spectrometry; Quantitative proteomics; Targeted quantitation

Contents

1. Introduction

While proteomics is contributing to a wide-range of scientific disciplines, probably no area is more critical than the discovery of novel diagnostic and therapeutic biomarkers. While discoveries in molecular biology help to unlock mysteries of cell function and behaviour, the discovery of clinically useful biomarkers would have a direct impact on the survival of thousands of patients and could mean the difference between choosing the correct or incorrect therapy in cases where immediate treatment is critical. One indisputable truth is the high standards that need to be achieved if a protein is to be useful as a biomarker. If a biomarker is defined as a feature that can be used to measure the presence and progress of a disease or the effects of treatment it must be able to be measured

☆ This paper was presented at Biomarker Discovery by Mass Spectrometry, Amsterdam, The Netherlands, 18–19 May 2006.

* Tel.: +1 301 846 7286; fax: +1 301 846 6037.
E-mail address: veenstra@ncifcrf.gov.

1570-0232/$ – see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jchromb.2006.09.004

reproducibly and also be specific to a disease or treatment. For example, while an increase in the levels of certain acute phase response proteins are used to indicate inflammation, they do not specify the exact cause of inflammation. Even the well known biomarker prostate-specific antigen (PSA) is not absolutely specific for prostate cancer, as other disorders such as benign prostatic hyperplasia, can result in an elevated PSA level [1]. Since a biomarker needs to be quantitated with high precision and accuracy it should be sufficiently abundant that it does not strain the limits of detection and quantitation available with today's assays or instrumentation. Finally, the test designed to detect the biomarker must possess high sensitivity (i.e. indicate a positive test for patients who have the disease) and specificity (i.e. indicate a negative test for patients without disease).

Probably no technology has spurred the fervor in discovery of new biomarkers than mass spectrometry (MS). The developments made in coupling protein and peptide fractionation techniques directly with state-of-the-art MS instrumentation has made it possible to identify thousands of proteins in complex biological samples [2]. This ability to obtain wide proteome coverage, however, has brought with it challenges in how to integrate this type of discovery science with basic research. The first challenge deals with the percentage of the proteome that we are presently able to characterize. Based on results from the human genome project, the human genome is anticipated to contain on the order of 20,000–25,000 open reading frames (Fig. 1A) [3]. Unfortunately the number of proteins within a complex proteome, from a biofluid for example, is unpredictable. Considering all of the possible post-transcriptional and post-translational events that may occur, any human proteome sample could easily contain upwards of 100,000 different protein species. The second challenge is that while discovery proteomics has focused considerable effort on developing methods to characterize thousands of proteins in biological samples, however, basic research continues to be dominated by scientists who focus on a single, or a very small number (i.e. 2–5), protein in any study. This disconnect is present in many aspects of biological research such as phosphorylation mapping, protein quantitation, and simple protein identification. It is very apparent in the field of biomarker discovery and validation. In the course of using MS, in particular, for the discovery of novel biomarkers hundreds of differences in the abundance of proteins between biofluids obtained from diseased and control patients can be observed, however it is currently only possible to graduate a small number of these "potential" markers into a validation phase.

The challenge in the next few years will be to find ways to bridge this divide between discovery-driven science and basic research. While improvements in technology will continue to benefit this progress, there are other study design and physiological barriers that may be more difficult to overcome. At a very fundamental level, reliable cohorts of samples that are indicative of the disease being study can be difficult to obtain. Unless a well thought out research study is designed in collaboration with a clinical center, very few groups are likely to hand over their "precious" clinical samples to a proteomics discovery laboratory. When dealing with tissue samples, biopsies require invasive procedures to obtain and are generally not collected in retrospective manner. There is no standardization in the collection of biofluid samples and the effects of processing and preparing serum and plasma are not well understood. With the ability of state-of-the-art mass spectrometers to identify low-abundance proteins in blood [4], we are only beginning to understand the overall effect of long-term storage and freeze/thaw cycles.

While many of these issues can be resolved by establishing standard operating procedures (SOPs), there are more ominous challenges. Let's consider a liver tumor that is secreting a highly specific biomarker into the circulation system. The concentration of this marker is very high in the immediate vicinity of the tumor. Most biomarker discovery efforts that analyze biofluids, however, scrutinize samples (such as serum and plasma) that are collected at the patient's elbow. This distance allows the biomarker to travel through thousands of miles of veins,

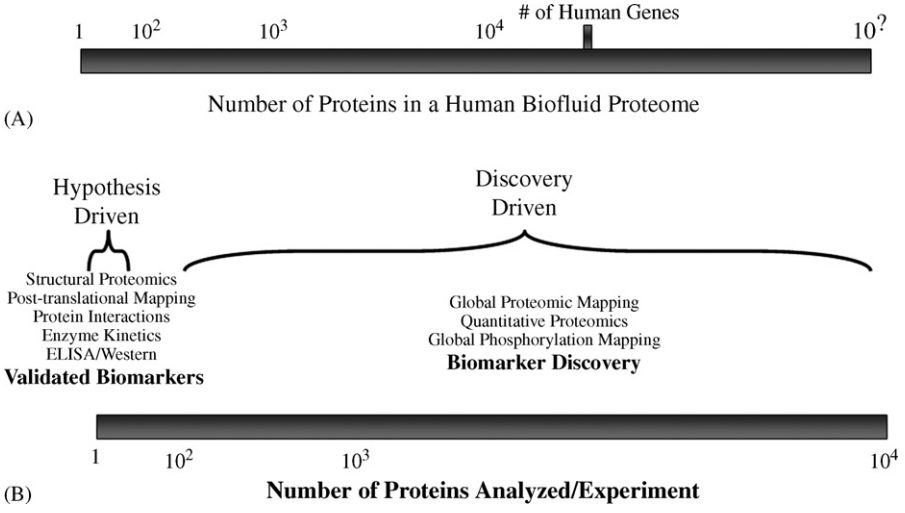


Fig. 1. (A) The disconnect between genomics and proteomics. While the number of genes within the human proteome will eventually be determined, it is impossible to accurately know the number of proteins within a complex proteome sample due such occurrences as post-transcriptional and translational modifications. (B) The disconnect between discovery-driven and basic research. While much of the focus of discovery-driven research is to acquire a low-density characterization of as many proteins in a proteome as possible, basic research is focused on acquiring a high-density characterization of a small number of proteins.

arteries, and capillaries in which it may be diluted to a vanishingly small concentration. Another physiological challenge involves the non-biased approach taken for biomarker discovery. On the surface it appears that most studies are trying to find the proverbial "needle-in-a-haystack". Unfortunately the situation is even direr than this analogy. In a typical study design in which vast numbers of proteins identified in biofluids collected from disease-affected patients are compared to matched controls, tens to hundreds of differences in protein abundances can be detected. The fundamental problem is that we lack the insight into which of these differences are related specifically to the condition being studied. Our inability to immediately recognize potential biomarkers that could be successfully validated essentially regulates these studies to finding a "needle-in-a-needlestack".

2. Quantitative strategies for biomarker detection

To identify novel diagnostic and therapeutic biomarkers, investigators focus on the discovery of proteins that are more or less abundant in samples obtained from patients with a specific disease compared to those acquired from healthy-matched control patients. There are a number of different MS-based methods for conducting such studies, and each has their particular advantages and disadvantages.

2.1. Two-dimensional polyacrylamide gel electrophoresis/mass spectrometry

Probably the most well known method of comparing protein abundances within complex proteomes is the combination of two-dimensional polyacrylamide gel electrophoresis (2D-PAGE) fractionation with MS protein identification (Fig. 2) [5,6]. In this technique, proteomes are extracted from two different samples that are being compared and then separated using 2D-PAGE (which separates proteins based on their isoelectric point in one dimension and molecular mass in the other). The gels are stained to visualize the resolved proteins and then spots that appear to be more abundant in one gel compared to the other are excised. An in-gel tryptic digestion of the gel spot is conducted and the protein is identified by MS analysis of the resultant peptides followed by bioinformatic analysis against the appropriate genomic or proteomic database. This method has been a "workhorse" in the field of comparative proteomics. It provides a direct method by which to visualize changes in proteins between complex proteome samples and is able to resolve thousands of proteins. Criticisms such as the inability of 2D-PAGE to resolve membrane proteins and its lack of reproducibility have been to some extent tempered by the development of better reagents, techniques, and gel alignment software. Unfortunately 2D-PAGE is still limited in sensitivity and dynamic range. The two most commonly used biofluids in biomarker discovery are serum and plasma. The content of both of these samples is dominated by a handful of proteins such as albumin and immunoglobulins [7]. Direct 2D-PAGE analysis of plasma and serum results in large smears of these proteins that mask lower abundance proteins [8]. Therefore, depletion of these high abundance proteins must be performed prior to the

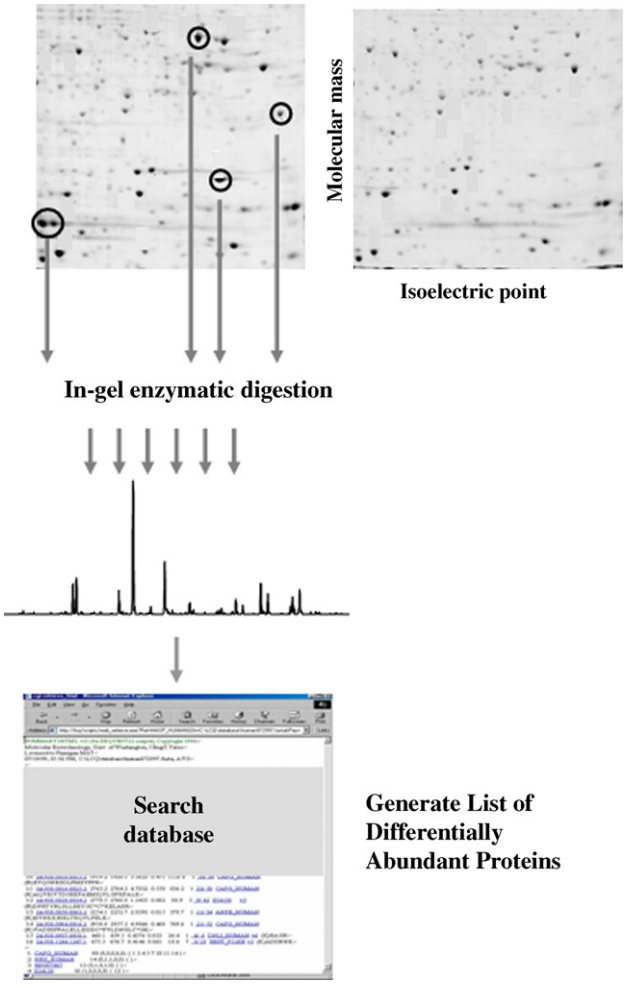


Fig. 2. Quantitative proteomics using two-dimensional polyacrylamide gel electrophoresis (2D-PAGE). In this method, comparative proteome samples are separated on distinct 2D-PAGE gels. After staining, protein spots that are more abundant on one gel compared to the other are excised from the gel. The protein(s) within the gel is then subjected to in-gel tryptic digestion and the resultant peptides are extracted and analyzed by mass spectrometry (MS). The MS data is then searched against the appropriate database to identify the protein(s) with the gel spot.

analysis of serum or plasma by 2D-PAGE. Its throughput is also comparatively slow making the comparison of multiple samples extremely time consuming. It does have the advantage, however, in that only those spots that appear differentially abundant need to be analyzed by MS.

2.2. Proteomic profiling

One method of biomarker discovery that generated great enthusiasm in the recent past is proteomic profiling (Fig. 3) [9,10]. In this method, a raw biofluid sample is applied to a chip containing spots made up of a specific chromatographic surface. Proteins within the samples are allowed to bind to the surface, which is then washing to remove non-binding species. The mass spectrum of the proteins bound to the chip spot is then recorded using a simple time-of-flight mass spectrometer. The mass spectra (referred to as a proteome pattern) of several (often

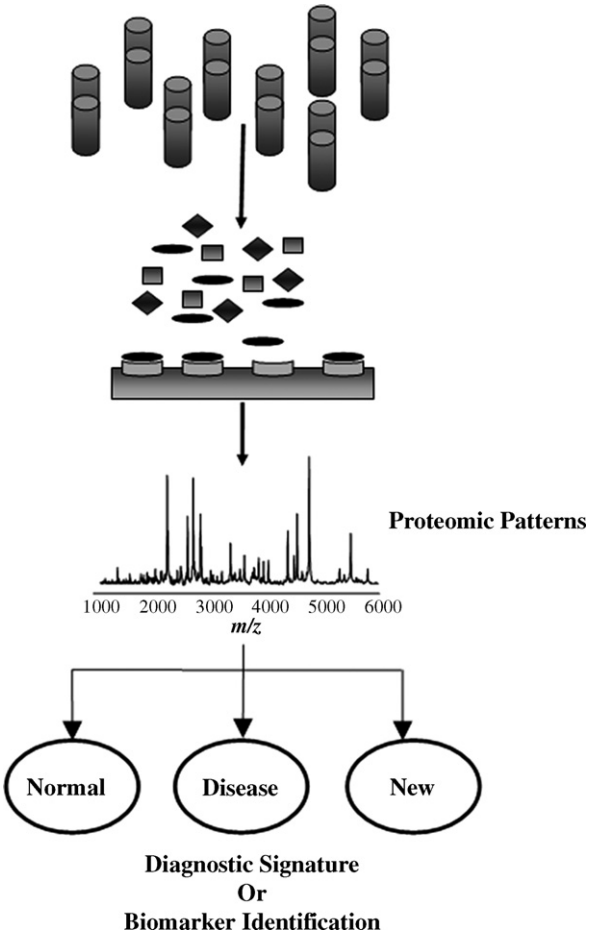


Fig. 3. Proteomic profiling using surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF/MS). In this method, the profile of proteins within a biofluid that are retained by a chromatographic surface is recorded by TOF/MS. Bioinformatic comparison of a series of spectra obtained of healthy and disease-affected individuals is used to classify the source of the sample.

hundreds) of samples obtained from disease-affected patients and healthy controls are acquired. The spectra obtained from disease-affected samples are then bioinformatically compared to those obtained of samples from the healthy controls. The bioinformatic algorithm attempts to classify the samples as coming from diseased or healthy patients (or from an unknown condition). Depending on the specific algorithm used, several peaks will be selected within the proteome pattern that allow for the correct classification of the sample source. These peaks can be either used as a direct means of diagnosis or can be identified using other proteomic methods. The major advantage of this method is that it is truly high-throughput and is capable of analyzing and comparing hundreds of biofluid samples in a matter of days. Many individual studies showed stunning results in the ability to correctly classify the sources of biofluid samples from either healthy or cancer-affected individuals.

Much of the early enthusiasm surrounding proteomic profiling has turned into notoriety [11]. Proteomic profiling has relied heavily on the use of peaks, not identified proteins, as the diagnostic determinants. Unfortunately, different algorithms would pick out different peaks as being diagnostic. Lab-to-lab reproducibility has yet to be demonstrated for this technique, as different labs analyzing the same disease states detect different diagnostic peaks in their analysis. In cases where the diagnostic peaks have been identified, they invariably turn out to be high abundant proteins (or fragments thereof) that are related to acute-phase response or inflammation, but lack disease-specificity. While this finding has been most commonly attributed to the insensitivity of the mass spectrometer used in these studies, it is actually related to its limited dynamic range. Essentially, the entire method uses a crude fractionation step and a single mass spectral acquisition to record as many species as possible that remain bound to the protein chip surface. This strategy limits the ability to detect low abundant proteins within a matrix of higher abundant blood proteins such as albumin. Until issues related to irreproducibility are resolved and this technology is shown capable of identifying a validated biomarker it is unlikely that it will regain any of the momentum it enjoyed previously.

2.3. Stable-isotope proteome tagging

A popular method of quantitatively comparing complex proteome samples, without the requirement of 2D-PAGE, is the use of stable-isotope tagging [12]. A popular method of stable-isotope tagging, isotope-coded affinity tags (ICATs) [13], which represents a good model for most of these types of studies, is shown in Fig. 4. The proteomes are extracted from two comparative samples and are then labeled with functionally and chemically identical reagents (in this case the ICAT reagents) that differ in their mass (i.e. 9.03 Da) based on their stable-isotope content (i.e. nine carbon-13 atoms in the heavy ICAT reagent in place of carbon-12 atoms in the light version). Once the proteins are differentially labeled, the two proteome samples are combined and digested into tryptic peptides. These peptides are then passed over an avidin column to extract out the stable-isotope tagged peptides. The ICAT reagent is unique in that it has iodoacetamide and biotin groups at opposite ends resulting in the modification of cysteinyl residues and the ability to reclaim these peptides using avidin chromatography. The biotin portion is then removed from the peptides and they are analyzed through a combination of multidimensional chromatography coupled directly on-line with data-dependent MS/MS. The mass spectrometer is operated in such a way that an MS scan is used to quantitate the relative abundance of the peptide within the different samples and MS/MS is used to identify the peptide in the same experiment. The net result is a list of identified proteins with a measure of their relative abundance between the samples being compared. Other stable-isotope labeling approaches both utilizing chemical modification and metabolic labeling have also been developed [4,14,15]. While slightly different than ICAT, they all use stable-isotopes and ultimately result in the same types of data sets.

Stable-isotope labeling methods have shown the capability of quantitating thousands of proteins in complex biological samples [16]. Unfortunately, they suffer similar disadvantages to 2D-PAGE. They are low throughput, requiring days to compare two samples. They are generally limited to comparing two samples; however, the development of iTRAQ has allowed

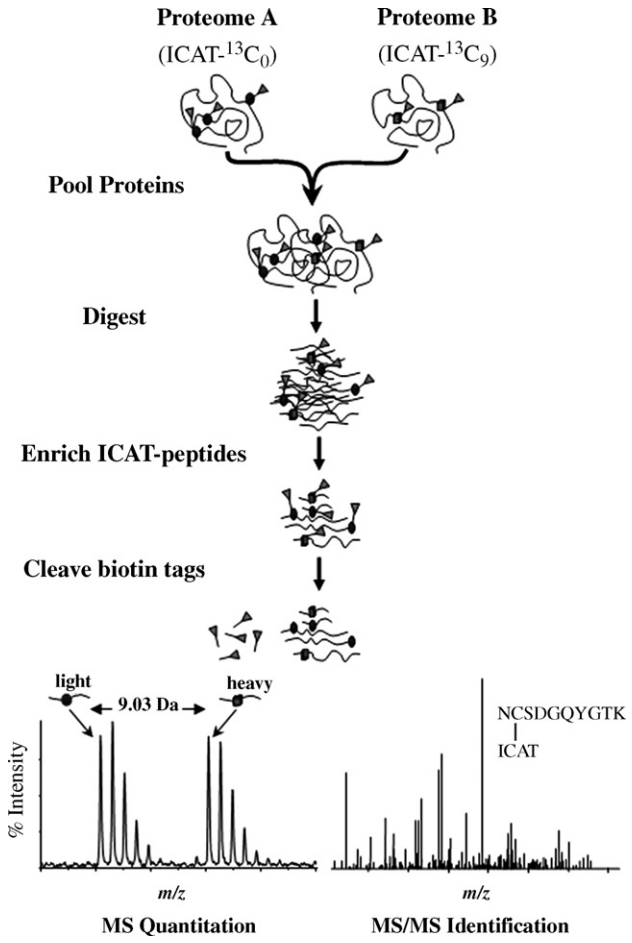
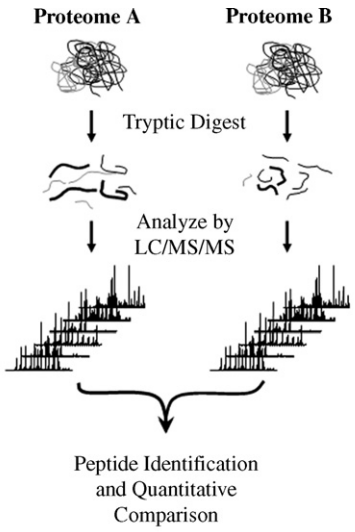


Fig. 4. Quantitative proteomics using isotope-coded affinity tags (ICAT). In this method, comparative proteome samples are labeled with chemically identical reagents that differ only by their carbon isotope content (i.e. nine carbon-12 atoms for the light reagent and nine carbon-13 atoms for the heavy reagent). After modification of the proteins, the proteome samples are combined and digested into tryptic peptides. The ICAT-modified peptides are extracted using avidin chromatography by virtue of the biotin moiety on the terminus of the ICAT reagents. After removal of the biotin portion, these ICAT-peptides are analyzed by reversed-phase liquid chromatography coupled directly on-line with a mass spectrometer. The mass spectrometer is operated in a data-dependent tandem mass spectrometry (MS/MS) mode, enabling the relative quantitation of the peptide in the two samples to be measured in the MS mode as well as its identity be discerned from the data acquired by MS/MS.

up to four samples to be compared simultaneously. Techniques that use metabolic stable-isotope labeling are impossible for the study of human samples. While they have made a major impact in the analysis of cellular and tissue proteomes, stable-isotope labeling methods, particularly ICAT, have not been widely used in biomarker discovery. The reasons for this are not readily obvious. It is possible that the domination of serum and plasma by a few high abundant proteins impacts the chemical labeling of lower abundant proteins by the stable-isotope reagents.

2.4. Subtractive proteomics

In an effort to simplify and increase the throughput of biomarker discovery many investigators are turning towards



$$\text{Protein Relative Abundance} = \frac{\#\text{ of Peptides Identified Proteome A}}{\#\text{ of Peptides Identified Proteome B}}$$

Fig. 5. Quantitative analysis using subtractive proteomics. In this method, proteome samples are digested into tryptic peptides and analyzed by liquid chromatography/tandem mass spectrometry (LC/MS/MS). The number of unique peptides identified for that specific protein is used as a measure of its relative abundance.

using a so-called subtractive proteomics approach [17]. This method does not use gels or stable-isotopes, but simply relies on quantitating proteins based on the number of peptides identified for each species (Fig. 5). In this method, the proteome is extracted from a series of biological samples and then digested into tryptic peptides. The tryptic peptides are then analyze using multidimensional chromatography coupled directly on-line with a mass spectrometer operating in a data-dependent MS/MS mode. The relative abundance of a protein is simply determined based on the number of peptides identified for that specific protein in the samples being compared. For example, if four peptides are identified for PSA in serum sample A and only one peptide is identified in serum sample B, the conclusion is made that PSA is four-fold more abundant in sample A compared to B.

The quantitative hypothesis is based on the fact that the number of unique tryptic peptides identified for a given protein is related to its abundance in the sample. The classic example of this is albumin in serum. If serum is simply digested into tryptic peptides and then analyzed by LC/MS/MS, a huge number of peptides from albumin will be readily identified while it is unlikely that a single peptide from a low-abundance protein such as cytokines will be observed. This result is directly related to the concentration of albumin (i.e. ~60 mg/mL) compared to cytokine proteins (i.e. in the ng/mL range). The denominator of this equation is the number of possible tryptic peptides that could be identified for a specific protein. This parameter is important as some proteins, particularly histones, may provide a large number of tryptic peptides, however, many of them may not be amenable to MS detection based on their size.

This subtractive approach is attractive for biomarker discovery mainly because of its inherent simplicity. Very little sample

preparation prior to MS analysis is required, save depletion of high abundance proteins. Thousands of peptides and proteins can be identified using this technology. An unlimited number of samples can be compared to one another. Like most techniques, however, it also has its disadvantages. It is relatively low-throughput. Each sample would take a minimum of 1 day to acquire the necessary data even if the whole process was automated. The quantitative comparison method is imprecise compared to stable-isotope labeling methods and therefore changes less than three-fold cannot be accurately determined with a high confidence level. Low abundance proteins, while detectable, may not provide enough unique peptide identifications to be quantitated using this method.

3. Bioinformatic analysis of quantitative proteomic data

3.1. Protein and peak identification

While a large number of peaks or identified proteins can be quantitatively measured using the approaches described above, the difficulty is how to turn this data into potentially validatable biomarkers. For those quantitative approaches that deal with identified proteins or peptides (i.e. 2D-PAGE, stable-isotope labeling, subtractive proteomics) there are a number of informatics solutions available for performing mass mapping and database searches using MS/MS spectra. A variety of algorithms including SEQUEST, Mascot, and ProteinProphet have been developed to search raw MS or tandem MS data against a variety of different protein databases [18]. Many times the choice of which algorithm to choose comes down to the personal preference of the laboratory. Two of the most common choices for a comprehensive multi-species non-redundant protein database are the mass spectrometry protein sequence database (MSDB) from the European Bioinformatics Institute [19], and the non-redundant database from the National Center for Biotechnology Information [20]. Both of these databases are comprised of unique composite protein sequences from multiple species produced from a number of source databases.

There are many issues related to the handling of non-gel based proteomic data. A major issue is simply the size of the data files, which can be ten gigabytes or greater for a single LC/MS/MS run. An efficiently operated mass spectrometer can produce upwards of one terabyte of data per year. Storage and analysis of these data sets can quickly overwhelm a stand alone PC, therefore most large MS-based proteomic laboratories have access to computer clusters and copious storage space. One simple solution to speed up protein identification is to filter spectra before they are analyzed. A typical LC/MS/MS experiment can generate on the order of 7000 tandem MS spectra per hour or which only a small percentage give rise to a useful identification. Filtering programs can be applied to the raw tandem MS data to remove noise and low quality spectra prior to peptide identification [21]. These algorithms typically enable the peptide identification analysis to be completed in half the time.

Proteomic pattern data is analyzed using different algorithms than for the other quantitative methods. There are essentially two different ways in which investigators have bioinformatically treated proteomic pattern data. One of these methods uses various software techniques such as support vector machines, decision trees, principle component analysis, and genetic algorithms to find peaks that enable spectrum from the two comparative sample sets to be segregated [22]. In most cases, peaks are not further identified and the collection of peaks are used as "biomarkers" for diagnosis. In the other approach, the intensities of the peaks within the spectra obtained from one class of samples is directly compared to those obtained from the other class to identify statistically significant changes in a peak's intensity between sample cohorts. Peaks that have been found to undergo a significant change in intensity are then subjected to identification through isolation of the responsible protein followed by peptide mapping or tandem MS [23].

3.2. Assessment of quantitative data for biomarker discovery

The sheer amount of quantitative data that can be acquired by MS-based methods enables large numbers of differences between comparative samples to be discovered. The difficulty is in establishing which differences are most important and likely to survive downstream pre-clinical validation. One constant in quantitative analysis of biofluids using tandem MS methods is the lack of throughput. To analyze what would be considered a small clinical cohort of samples (i.e. between 75 and 100) requires months of effort. Therefore, it is difficult to conduct repeat analyses of samples as is done in array experiments. Validation studies need to be conducted against specific proteins using higher throughput methods such as Western blotting or ELISA. But which proteins should be graduated to pre-clinical validation? Many differences, such as inflammatory or acute-phase response proteins, can be ruled out as potential biomarkers since they do not possess disease specificity. Quantitative changes in the proteome are often compared to those observed in an array experiment, however, numerous studies have shown only modest correlation between the amount of a protein and its transcript's abundance [24] and these types of comparisons within biofluids are rare. The hurdle in determining the significance of any observed quantitative change in large proteomic datasets has been the single greatest barrier in the discovery of biomarkers.

The stringency level of the discovery phase dictates the number of biomarker candidates that graduate to pre-clinical validation phase [25]. As mentioned above, LC/MS/MS analyses measures thousands of analytes but is limited to at most tens of comparative samples. It is therefore only possible to conduct validation studies on a small subset of identified proteomic differences. These candidate proteins are often selected based on biological knowledge, quality of the quantitative MS data, and the availability of reagents (e.g. antibodies) to confirm its potential as a biomarker. This "hit-or-miss" strategy is sub-optimal and illustrates the need to compliment MS-based with other experimental observation in order to increase the chance of a potential biomarker progressing to a validated biomarker.

4. Targeted approaches to quantitate biomarkers

Most of this review has been focused on un-biased methods to attempt to find novel biomarkers that can be used for diagnosis or therapeutic monitoring. The general thought is that MS will play a major role in discovery; however, the validation and routine monitoring of biomarkers will be accomplished through the development of affinity reagents such as antibodies. The production of a highly specific, high affinity antibody to a newly discovered protein biomarker, however, is never a certainty and can take a considerable amount of time, at a great expense, to produce. There are many analytical issues related to antibody detection, including non-specificity, cross-reactivity, and lot-to-lot variation. While it is inarguable that antibodies will continue to play a major role in biomarker detection, the question needs to be considered if MS can play a role beyond discovery.

What if MS identifies a novel biomarker, or a useful biomarker is presently known, but no useful affinity-based test is available? While much of the focus has been on utilizing the attributes of MS, such as high sensitivity, resolution, and mass accuracy, for identifying thousands of species, it is sometimes forgotten that these characteristics also enable this technology to quantitate a single component in a complex mixture. This ability has been very effectively demonstrated in a clinical setting through the detection of in-born errors of metabolism [26]. This MS/MS-based technology provides a multianalyte metabolic profile of blood samples obtained from newborns and has been used to detect diseases such as phenylketonuria and disorders arising from errors in fatty acid oxidation and organic acid metabolism. This test has worked very effectively for monitoring metabolites; therefore, similar MS-based methods should also be applicable for monitoring protein biomarkers.

The ability to monitor a specific protein analyte in a complex proteome sample is conducted through methods such as selected reaction monitoring (SRM) (Fig. 6). In this method, LC is used to fractionate a proteome mixture prior to its infusion into the mass spectrometer. The illustration in Fig. 6 shows how the ions are manipulated when using a triple quadrupole mass spectrometer. In an SRM analysis, the elution time of the peptide is generally known. Therefore, at a specific time during the LC separation the instrument can be instructed to isolate a specific mass-to-charge (m/z) value within the first quadrupole (Q1) region of the instrument. This peptide ion then enters the collision cell (Q2) and is fragmented. Specific fragments are then isolated in the Q3 region and pass through to the detector. Why are fragments scanned instead of just monitoring the intact peptide? In two words: specificity and sensitivity. Although a known m/z value can be isolated during a specific time during the separation, it must be kept in mind that a human proteome samples is very complex and the degeneracy of peptide masses is very high.

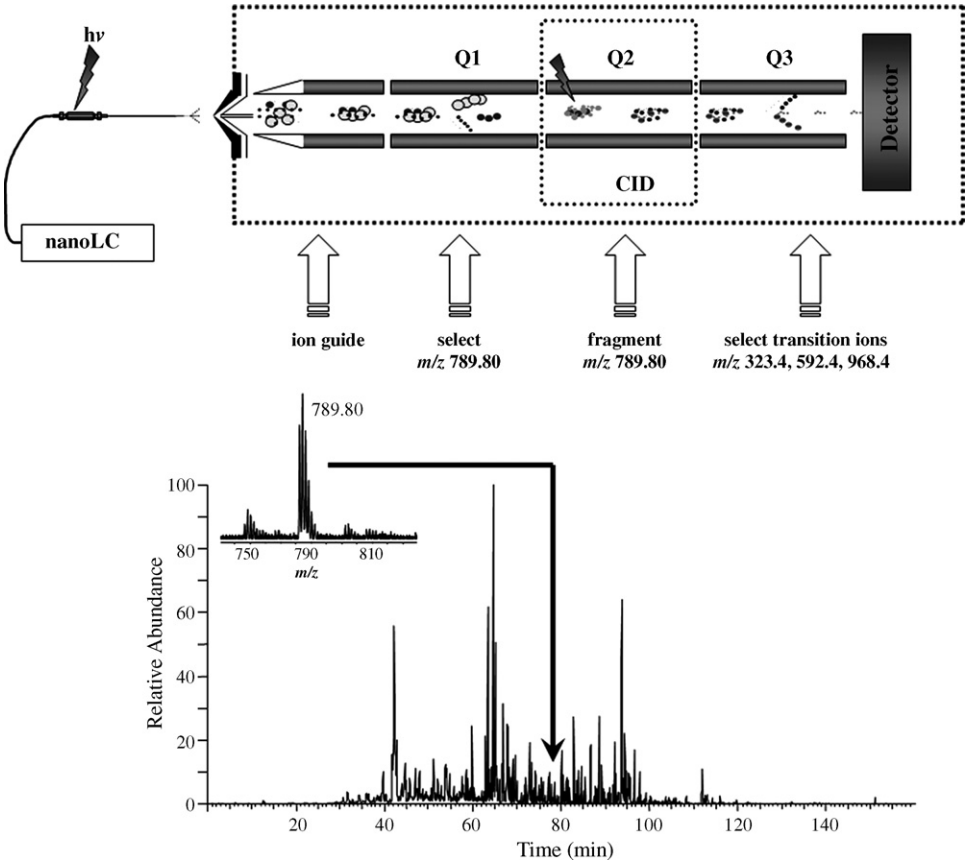


Fig. 6. Example of selected reaction monitoring (SRM) for biomarker quantitation. In this method, the first quadrupole (Q1) within a triple quadrupole mass spectrometer is used to isolate a specific peptide ion (i.e. m/z 789.80) that elutes at a specific time in a liquid chromatography separation of a complex mixture. This selected ion is fragmented within Q2 and specific fragment ions are allowed to pass through Q3 onto the detector.

Therefore, if there is any irreproducibility in the LC separation or inaccuracy in the mass measurement, the incorrect peptide could end up being selected. The fragments, however, produced from a peptide are very unique since they are a function of its amino acid sequence. Sensitivity is not only a function of signal intensity but also of the noise. In monitoring for fragment ions, a considerable amount of noise is excluded, thereby increasing the sensitivity of the measurement.

Since a vast majority of proteome studies conducted by MS examine tryptic peptides instead of intact proteins, it is important to determine which peptide functions best as a surrogate for its protein of origin. While some empirical rules can be established (e.g. no residues within the peptide should be susceptible to modification such as oxidation), the optimal peptide surrogate needs to be experimentally determined. This determination can be accomplished by analyzing a tryptic digest of the protein biomarker and finding those peptides that provide an intense signal, give reproducible fragment ions, and are unique to the protein of interest. Once determined, a stable-isotope version of this peptide is synthesized and used as an internal standard to measure the absolute quantity of the protein of interest in a biological sample.

One of the first examples of using multiplexed SRM to monitor proteins in a biological sample was recently published by Anderson and Hunter [27]. In this study, multiple reaction monitoring (MRM) was used to assay 53 high and medium abundance proteins in human plasma. They found that 47 of these assays produced quantitative data with coefficient of variations (CVs) ($n = 10$) of 2–22%. Peptides from proteins such as L-selectin could be reliably measured, showing that proteins in the μg/mL concentration level can be reliably quantitated in plasma. While immuno-depletion of six high abundance proteins significantly improved CVs compared with whole plasma, the targeted analytes could be detected in both sample types. Studies conducted in our laboratory, however, have shown that tailoring the sample preparation to a desired biomarker can enable the routine detection into the attomole concentration level (unpublished results).

5. Conclusions

The advances made in proteomic technology, primarily in the field of MS, have equipped us with the ability to scrutinize proteome samples to a far greater extent then ever possible. As described in this article, there are many options available for measuring the relative abundances of proteins in clinical samples. Unfortunately the number of biomarkers that have ultimately been successfully validated using these discovery approaches is discouraging. The fault for this fact, however, does not rest solely on the technology: there are a number of physiological characteristics of biofluids that makes the challenge very difficult. In fact, MS-based studies are able to come up with very large numbers of "potential" biomarkers. The challenge is how to identify those that have the highest chance of being validated in a well-controlled clinical trial. Validation of a single biomarker is expensive in terms of money and time. Therefore, it is impossible to graduate a large number of potential biomarkers to a validation phase. Unfortunately it is difficult to inherently recognize those proteins identified in the discovery phase that may turn out to be the best diagnostic or therapeutic biomarker. While the situation seems bleak, encouragement can be found in the progress that has been made in the past 5 years that has allowed investigators to even attempt the types of biomarker studies that are being conducted today.

Most of the biomarker discovery efforts being conducted using MS focus on identifying wild-type proteins that are simply more or less abundant in a diseased condition when compared to the healthy state. This paradigm fits well with the existing paradigm for markers such as PSA and cancer antigen-125, in which the concentration of a protein within a clinical sample is used to make a diagnostic decision. What about proteins that are not predicted by any known genome or proteome sequence? Unfortunately, we are often hampered in the analysis of MS data by our inability to discovery aberrant "unpredictable" proteins. One example of an aberrant protein acting as a disease biomarker is illustrated in the identification of antiproliferative factor (APF), an indicator of interstitial cystitis [28]. Interstitial cystitis is a painful bladder disorder, which is characterized by a frequent urgency to urinate and a thinning of the bladder epithelial lining. The biomarker for this disorder, APF, was recently identified as a glycosylated nine-residue peptide. The modification and peptide sequence of this marker was such that it would have been unlikely to be found in a typical experiment in which proteomes are enzymatically digested and analyzed by MS. Only through the inclusion of other bio-analytical tools was this molecule recognized and eventually identified by MS. Considering the cellular phenotypes observed in many diseases, in particular cancers, it may not be surprising that other atypical proteins are identified in the future as being useful biomarkers for such conditions.

Acknowledgements

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.

References

[1] F.H. Schroder, Can. J. Urol. 12 (2005) 2.
[2] T. Kislinger, A.O. Gramolini, D.H. MacLennan, A. Emili, J. Am. Soc. Mass Spectrom. 16 (2005) 1207.
[3] http://www.ornl.gov/sci/techresources/Human_Genome/home.shtml.
[4] B.L. Hood, M. Zhou, K.C. Chan, D.A. Lucas, G.J. Kim, H.J. Issaq, T.D. Veenstra, T.P. Conrads, J. Proteome Res. 4 (2005) 1561.
[5] M.C. Pietrogrande, N. Marchetti, F. Dondi, P.G. Righetti, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 833 (2006) 51.
[6] P. Weingarten, P. Lutter, A. Wattenberg, M. Blueggel, S. Bailey, J. Klose, H.E. Meyer, C. Huels, Methods Mol. Med. 109 (2005) 155.
[7] L. Anderson, J. Physiol. 563 (2005) 23.
[8] E.B. Altintas, A. Denizli, J. Chromatogr. B Analyt. Technol. Biomed. Life Sci. 832 (2006) 216.

[9] V. Seibert, M.P. Ebert, T. Buschmann, Brief. Funct. Genomic. Proteomic 4 (2005) 16.
[10] E.F. Petricoin, A.M. Ardekani, B.A. Hitt, P.J. Levine, V.A. Fusaro, S.M. Steinberg, G.B. Mills, C. Simone, D.A. Fishman, E.C. Kohn, L.A. Liotta, Lancet 359 (2002) 572.
[11] M. Zhou, T.P. Conrads, T.D. Veenstra, Brief. Funct. Genomic. Proteomic 4 (2005) 69.
[12] S.E. Ong, M. Mann, Nat. Chem. Biol. 1 (2005) 252.
[13] S.P. Gygi, B. Rist, S.A. Gerber, F. Turecek, M.H. Gelb, R. Aebersold, Nat. Biotechnol. 17 (1999) 994.
[14] S.E. Ong, L.J. Foster, M. Mann, Methods 29 (2003) 124.
[15] M. Heller, H. Mattou, C. Manzel, X. Yao, J. Am. Soc. Mass Spectrom. 14 (2003) 104.
[16] K.A. Conrads, M. Yi, K.A. Simpson, D.A. Lucas, C.E. Camalier, L.R. Yu, T.D. Veenstra, R.M. Stephens, T.P. Conrads, G.R. Beck Jr., Mol. Cell. Proteomics 4 (2005) 1284.
[17] E.C. Schirmer, L. Florens, T. Guan, J.R. Yates 3rd, L. Gerace, Science 301 (2003) 1380.
[18] E. Kolker, R. Higdon, J.M. Hogan, Trends Microbiol. 14 (2006) 229.
[19] http://csc-fserve.hh.med.ic.ac.uk/msdb.html.
[20] http://www.ncbi.nlm.nih.gov/RefSeq/.
[21] W. Sun, F. Li, J. Wang, D. Zheng, Y. Gao, Mol. Cell. Proteomics 3 (2004) 1194.
[22] Y. Liu, Technol. Cancer Res. Treat. 5 (2006) 61.
[23] N. Escher, B. Spies-Weisshart, M. Kaatz, C. Melle, A. Bleul, D. Driesch, U. Wollina, F. von Eggeling, Eur. J. Cancer 42 (2006) 249.
[24] C.J. Hack, Brief. Funct. Genomic. Proteomic 3 (2004) 212.
[25] N. Rifai, M.A. Gillette, S.A. Carr, Nat. Biotechnol. 24 (2006) 971.
[26] D.H. Chace, T.A. Kalas, Clin. Biochem. 38 (2005) 296.
[27] L. Anderson, C.L. Hunter, Mol. Cell. Proteomics 5 (2006) 573.
[28] S. Keay, Z. Szekely, T.P. Conrads, T.D. Veenstra, J. Barchi, C. Zhang, K. Koch, C. Michjeda, Proc. Natl. Acad. Sci. U.S.A. 101 (2004) 11803.