A Top-Down Proteomic Approach for the Discovery of Novel Serum Biomarkers of Pregnancy-Related Disease

Karen Merrell
Brigham Young University - Provo

A TOP-DOWN PROTEOMIC APPROACH FOR THE DISCOVERY OF NOVEL SERUM BIOMARKERS OF PREGNANCY-RELATED DISEASE

By

Karen Merrell

A dissertation submitted to the faculty of

Brigham Young University

In partial fulfillment of the requirements for the degree of

Doctor of Philosophy

Department of Chemistry and Biochemistry

Brigham Young University

December 2009

BRIGHAM YOUNG UNIVERSITY

GRADUATE COMMITTEE APPROVAL

of a dissertation submitted by

Karen Merrell

This dissertation has been read by each member of the following graduate committee and by majority vote has been found to be satisfactory.

Date	Steven W. Graves, Chair
Date	Paul B. Farnsworth
Date	Craig D. Thulin
Date	Gerald D. Watt
Date	Barry M. Willardson

BRIGHAM YOUNG UNIVERSITY

As chair of the candidate's graduate committee, I have read the dissertation of Karen Merrell in its final form and have found that (1) its format, citations, and bibliographical style are consistent and acceptable and fulfill university and department style requirements; (2) its illustrative materials including figures, tables, and charts are in place; and (3) the final manuscript is satisfactory to the graduate committee and is ready for submission to the university library.

Date

Steven W. Graves
Chair, Graduate Committee

Accepted for the Department

Date

David V. Dearden
Graduate Coordintor,
Department of Chemistry and Biochemistry

Accepted for the College

Date

Thomas W. Sederberg
Associate Dean,
College of Physical and Mathematical Sciences

Abstract

A TOP-DOWN PROTEOMIC APPROACH FOR THE DISCOVERY OF NOVEL SERUM BIOMARKERS OF PREGNANCY-RELATED DISEASE

Karen Merrell

Department of Chemistry and Biochemistry

Doctor of Philosophy

The serum fraction of blood is an ideal material in which to search for novel biomarkers for disease. It is easily obtained through relatively non-invasive means, routinely collected, and a rich treasure-trove of information about the health of an individual. Cells react to signal molecules, take up nutrients, and release waste products, fragments that are the result of proteolysis, and other molecules out into the bloodstream. If these components are unique to the cells in question, that part of the complex mixture that is the blood stream can potentially characterize the health of the tissue or organ those cells are a part of.

Serum is dense with proteins that span over ten orders of magnitude in size and abundance. The top 22 most abundant proteins in serum account for 99% of the total protein. These abundant proteins are well-characterized and not useful in a search for novel biomarkers for disease. Removal of these large proteins is accomplished using an organic-solvent precipitation step. Analyzing the resultant mixture of low-molecular-weight serum peptides using cLC-MS produces large, data-rich, and very complex data files. We have developed a manual analysis method we have developed that is capable of performing all of the processing steps necessary to identify novel biomarkers for disease as well as a method for the sequencing of low-abundance, highly charged peptide species without additional sample preparation.

These methods are applied to two serum sample sets collected to investigate two pregnancy-related diseases: preterm birth, and preeclampsia. Three novel biomarkers of preterm birth have been identified and a combination of these with 5 previously studied markers can predict women who will have preterm birth with a sensitivity of 89.9% and a specificity of 81.0%. Nineteen different molecular species have been identified that predict women at risk for preeclampsia with a p-value of <0.05. Weighted combinations of various groups of the 19 biomarkers can increase the sensitivity up to 96% and the specificity up to 100%.

The use of cLC-MS in the search for novel serum biomarkers of pregnancy-related disease allows for seamless integration from potential biomarker selection to polypeptide sequence identification.

Acknowledgements

I owe my success to all of the wonderful people who have supported, assisted, and loved me during my studies at BYU. I would like to thank my mentors Dr. Steven Graves, Dr. Craig Thulin, and Dr. M. Sean Esplin for your support, patience, love of science, sense of humor, and the multitude of things you have taught me. I thank my co-workers and friends Sarah, Mike, Chen, Brad, Ryan G., Ryan H., Wayne, Aaron, Tom, Moana, and for the many people who have directly contributed to this work: Katie, Nate, Peter, Dave, Esther, Jun, Emé, Lilly, Tanielle, Brandon, Rachel, Mitch, Dan, Brent, and Caroline. I would especially like to thank my family for being the rock of support that I have built my life upon. Your unconditional love, constant encouragement, and sacrifices on my behalf are more than I could have asked for. I love you!

Supported by grants from the National Institute of Child Health and Human Development (HD21410, HD21414, HD27860, HD27861, HD27869, HD27883, HD27889, HD27905, HD27915, HD27917, HD19897 and HD36801).

Table of Contents

List of Tables

List of Figures

Abbreviations

Basic local alignment search tool – BLAST

Capillary liquid chromatography – cLC

Collision energy – CE

Collisionally-induced dissociation – CID

Electron capture dissociation – ECD

Fourier-transform ion cyclotron resonance – FTICR

Independent Data Acquisition – IDA

Inter-alpha-trypsin inhibitor heavy chain 4 – ITIH4

Liquid chromatography mass spectrometry – LCMS

Low molecular weight – LMW

Mass spectrometry – MS

Multi-channel Acquisition – MCA

Matrix-assisted laser desorption ionization – MALDI

Nominal molecular weight limit – NMWL

One-dimensional polyacrylamide gel electrophoresis – PAGE

Prostate Specific Antigen – PSA

Quadrupole – Q

Rheumatoid Arthritis – RA

Spontaneous preterm birth – SPTB

Surface-enhanced laser desorption ionization – SELDI

Tandem mass spectrometry – MS/MS

Time-of-flight – TOF

Two-dimension polyacrylamide gel electrophoresis – 2D-GE

Chapter 1 – Introduction

Protein Analysis

Frederick Banting John Macleod

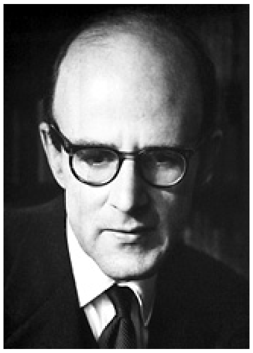
(http://nobelprize.org)

Prior to development of modern instrumentation, different approaches to protein analysis were used to investigate the molecular properties of individual proteins. Some of the properties commonly investigated were biological activity, three dimensional structure, and amino acid sequence. Though somewhat limited in their approaches, these studies provided valuable insight that advanced the understanding of the importance of proteins. This is highlighted by some of the Nobel prizes awarded for such ground breaking research in these areas. In 1923 the Nobel Prize in medicine was awarded to Frederick Banting and John Macleod for the discovery of insulin. They isolated a protein from pancreatic extracts, and demonstrated that the protein retained biological activity. This protein kept a pancreatectomized dog alive all summer because the extract lowered the level of sugar in the blood. (Bliss 1982) Fred Sanger won a Nobel Prize in chemistry in 1958 for his work on the structure of proteins, especially that of insulin. In that same year, John Kendrew and associates successfully determined the structure of myoglobin by high-resolution X-ray crystallography. (Kendrew 1958) For their studies of the structures of globular proteins, John Kendrew shared the 1962 Nobel Prize in chemistry with Max Perutz. (nobelprize.org) Edman sequencing has been a standard

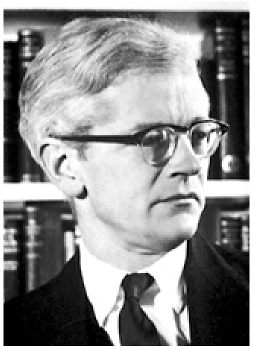
Frederick Sanger

(http://nobelprize.org)

method for protein amino acid sequence identification since it was developed in 1970. (Edman 1970) This method employs a stepwise chemical degradation of a protein from its N-terminus. Because N-terminal modifications block sequencing using this method, a method employing enzymatic digestion step prior to Edman sequencing of internal peptides was developed. (Abersold 1987) Sequencing of enzymatically digested proteins, the online publication of genomic databases, and the translated protein databases rendered from them prepared the way for mass spectrometry (MS) based methods. (Patterson 2001)



Max Perutz



John Kendrew

(http://nobelprize.org)

Nearly all of the processes in living cells are accomplished by the effects of proteins. The genome regulates the synthesis of all of these proteins. Many genome projects have sequenced the structure and composition of entire genomes for many different organisms. For example, there are ~8,000 genes in nematodes compared with the ~22,000 genes in humans. The difference in organism complexity between nematodes and humans cannot be explained by the genomes themselves: That of humans is not that much larger than that of nematodes. The complexity of the human organism compared to that of the nematodes must then be explained by the products of the genomic “blueprint” rather than the blueprint itself. (Shaw 2008) Because the genomes of many organisms have already been sequenced, it is often possible to predict protein sequences that may result from the translation of specific gene sequences. There is no way to know for sure which of these proteins are expressed *in vivo*. Analysis of the rough draft of the human genome showed that there were fewer protein-coding genes in the human genome than there are proteins in the human proteome (~22,000 genes versus ~400,000 proteins). (Yim 2006) This increase in protein diversity is thought to be a result of alternative splicing, differential proteolysis, and posttranslational modification of proteins. The final state of these proteins cannot be predicted through analysis of the genome. Thus, to characterize the protein complement of cells and tissues of interest, a proteomics-based approach is needed. (Zubrig 2008 Peptidomics)

Proteomics

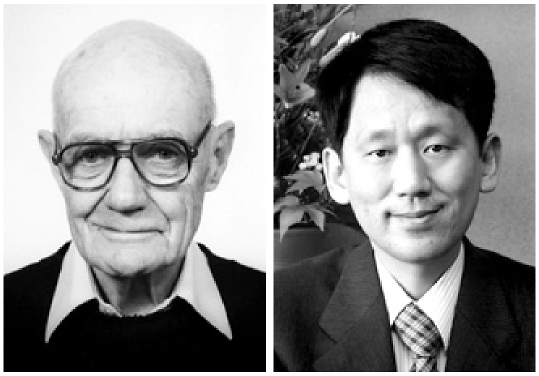
The term proteome was coined in 1995 and defined as "the total protein content of a genome." (Kahn 1995, Wasinger 1995) This term was coined to make an analogy with genomics, and is generally regarded as the "next step" in biological discovery. (Zubrig 2008 Peptidomics) Proteomics is the study of the proteome and involves the study of the protein complement and their interactions in a cell, organ, tissue, or entire organism. (Lundblad 2006) The genome is a constant entity whereas the proteome is constantly changing and adapting as it interacts with the genome and the biochemical environment. (Zubrig 2008 Biomarker) Because the protein complement of even a single cell is varied, dynamic and complex, those researchers undertaking proteomic investigations have developed many sophisticated techniques. These techniques allow researchers to simplify and observe a representative segment or subset of the proteomes of biological materials or specimens. Current methods of sample protein separation include one-dimensional polyacrylamide gel electrophoresis (PAGE) to separate proteins by size, two-dimensional PAGE (2D-GE) to separate proteins by isoelectric point as well as size, and liquid chromatography to separate proteins by charge, size, or hydrophobicity.

PAGE was reported as early as 1970 (Laemmli 1970), and 2D-GE was reported five years later. (O'Farrell 1975) The 2D gel separating and blotting techniques are still widely and successfully used today despite limitations in reproducibility, sensitivity, and lengthy analysis time.

There seems to be three general areas of inquiry for proteomics. The first is analytical proteomics where the elucidation of the proteome is undertaken by analytical means, including microarrays, mass spectrometry and various separation techniques. The second is expression proteomics where some kind of stimulus is applied to a system and the resulting change in protein expression and the underlying pathways are observed. The third is biomarker identification where the protein complement from a healthy organism is compared with that from one with a certain disease. Biochemical differences in the proteome between cases and control can be utilized as biomarkers. (Lundblad 2006) Proteomics as an analytical tool has the potential to identify within a very complex mixture of thousands of proteins, those individual proteins associated with change in phenotype due to disease. (Anderson 1998) Subsequent studies can then be conducted to see whether or not the observed correlation can provide clues as to the cause of or the sequence of pathways leading to that particular disease.

Mass Spectrometry

With the advent of soft-ionization techniques like matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), mass spectrometry (MS) has been useable for the analysis of larger biomolecules. These breakthroughs have enabled the field of proteomics to expand beyond the mass and detection limits of 2D-GE. MS instruments consist of at least three basic components; an ionization module, a mass analyzer, and a detector. For ions to be able to move through the electromagnetic fields of the mass analyzer they need to be charged. The ionization module of instruments designed for the analysis of biological specimens typically adds protons to the molecules to be analyzed, thus conferring the charge necessary for analysis. The movement of these ions through the mass spectrometer is dependent on both mass and charge and is typically measured as mass over charge (*m/z*) ratio. Because the direct mass is not measured, the real mass of the ions can only be determined if the charge state can be determined. Determination of charge state is possible because of naturally occurring isotopes, but becomes more difficult with increasing size and charge states. (Zubrig 2008 Peptidomics) The use of a reflectron, also known as an electronic or ion mirror, is used in TOF MS instruments to reflect ions and substantially increases resolution and the ability to determine the mass of larger and more highly charged molecules. (Hortin 2006) The importance of MS analysis of biological molecules methods was highlighted when Koichi Tanaka and John Fenn were awarded the Nobel Prize for chemistry in 2002 for the invention of MALDI and ESI respectively.



John B. Fenn Koichi Tanaka

(http://nobelprize.org)

The most common types of mass spectrometers used for proteomic analysis are quadrupole (Q), ion-trap, time-of-flight (TOF), and Fourier-transform-ion cyclotron resonance (FT-ICR) instruments or their combinations (e.g., hybrid instruments such as Q-TOF, combining quadrupole and time-of-flight detectors). All mass spectrometers measure *m/z*, and so they all require that species to be analyzed carry a charge. Ionization of biomolecules can be achieved using either MALDI or ESI. With MALDI, sample is co-precipitated with a matrix onto a sample plate. Ionization is achieved by pulsing a laser at the co-precipitated sample. The matrix absorbs the laser energy and desorbs from the plate, taking the charged protein sample with it. Molecules ionized using this method carry single charges. With ESI, charged droplets are sprayed from a high voltage needle. A stream of nitrogen gas evaporates the solvent, leaving the singly or multiply-charged proteins in the gas phase. (Figure 1.1) (Zubrig 2008 Peptidomics)

High resolution mass spectrometers allow for accurate reading of fragmentation data caused by collision of the peptides of interest with inert gas, typically called peptide mass fingerprinting. This fragmentation mainly occurs at peptide bonds and yields a series of overlapping fragment ions that include either the C- or N- terminus of the peptide. The *m/z* differences between fragments differ by the exact weight of an amino acid and permits deduction of the amino acid sequence of that peptide. (Papayannopoulos 1995) This collision-induced method of fragmentation is referred to as tandem MS or MS/MS. The traditional proteomics role of MS/MS was limited to identifying the proteins in spots or bands from one- or two- dimensional polyacrylamide gels. 2D-GE has numerous limitations as a method for visualizing and/or resolving many types of proteins, including small proteins, hydrophobic proteins and even some soluble proteins. (Cutillas 2008)

MS has revolutionized protein identification by pushing sensitivity up by several orders of magnitude over previous detection and sequencing methods, and has shortened the identification process from many hours to just a few minutes. Using ESI MS/MS methods, identification of several hundred peptides per day can be achieved. The purity and amino acid composition of the peptides being analyzed determines their ionization efficiency. Some combinations of amino acids don't ionize well, and so as the mixture of peptides in a given sample competes for ionization, the signal of some peptides will be suppressed. Sensitivity also decreases as the complexity of the sample increases. If a

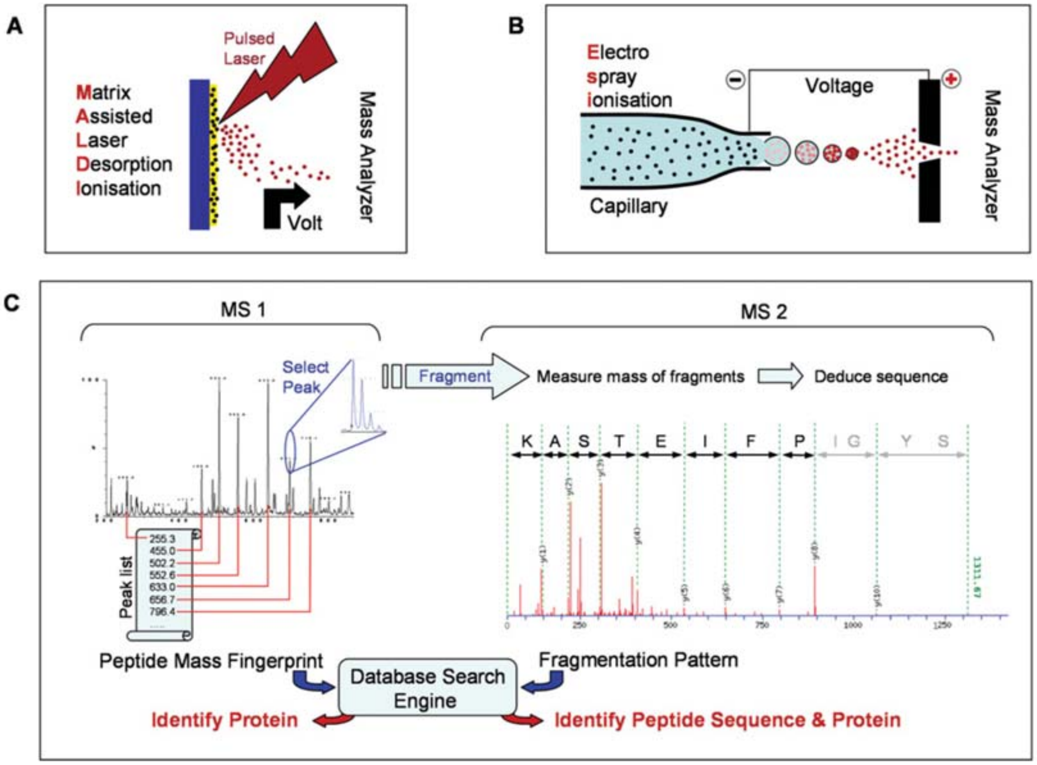


Figure 1.1. Principles of MS – MALDI (A) and ESI (B) Ionization methods for peptides. (C) A typical MS experiment. The left-hand panel shows a peptide mass fingerprint. A survey scan measures the peptide masses, which can then be inspected for peaks of interest. The right-hand panel demonstrates a determination of peptide sequence. Using MS, the peptide peak to be sequenced is selected from the survey scan and fragmented, usually by collision with gas. The fragment masses are determined by MS/MS. The peptide sequence can be deduced from the fragment masses. Because the fragmentation spectra rarely represent fragmentation between every peptide bond, de novo protein sequencing from the spectra alone is often not possible. Therefore, rather than trying to read the sequence from the mass spectrum, the fragmentation patterns are usually submitted to a database search engine for identification. This search compares the observed fragmentation pattern with theoretically expected peptide fragmentation patterns of the proteins predicted from the compiled genome databases. (Zübrig 2008 Peptidomics)

sample is dominated by high abundance peptides, it is difficult to detect the lower abundance peptides. This is a situation typical for serum or plasma samples where the protein concentration varies over more than ten orders of magnitude. Additionally, both sensitivity and mass accuracy drop off sharply with increasing size, so the most effective mass window is between 1 and 5 kDa. Because of this, proteins are usually digested into peptides prior to analysis with MS. (Zubrig 2008 Peptidomics)

In order to maximize the number of molecules analyzed in a complex sample, capillary liquid chromatography (cLC) is often coupled to ESI MS. This provides a powerful separation step that allows separation of analytes and increases sensitivity. (Abersold 2003, Issaq 2002, Zurbig 2006) Using cLC-MS, information on hundreds of polypeptides from an individual sample can be obtained relatively quickly. There is a great potential for finding novel biomarkers within this information. The physiological role of any potential biomarkers found using cLC-MS remains unknown unless their amino acid sequence can be determined. Because these biomarkers are most likely low molecular weight (LMW) and present in relatively low quantities in the serum, they cannot be easily isolated or purified. Potential biomarkers may even be fragments of larger proteins, representing both those that are proteolytically cleaved as well as those that are post-translationally modified. (Zubrig 2008 Peptidomics) Identification of such biomarkers requires *de novo* sequencing. A semiquantitative measurement of two different samples can be made by measuring the peptide signal intensity. This can be done without using labels as long as the experimental conditions are constant and the sample analysis order is randomized. (Radulovic 2004, Johansson 2006, Sköld 2008)

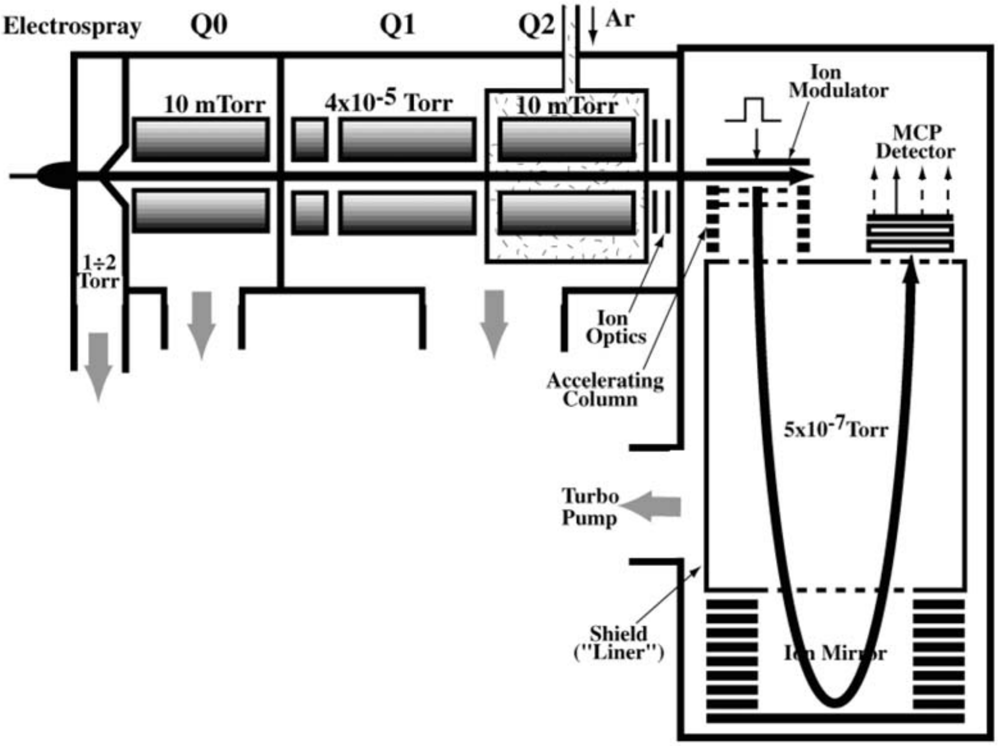


Figure 1.2. Schematic of Qstar QqTOF Mass Spectrometer. Q0 and Q2 are rf-only quadupoles used to focus ions into the next part of the instrument. Q1 may be operated in the rf-mode only (MS) or in the mass selective mode (CID MS/MS). Mass resolution is 1000-2000. TOF analyzer with ion mirror has a Mass resolution ~10,000. (from www.iccs.edu/QSTAR w.pdf)

Serum Proteomics for Biomarker Discovery

There are two main challenges in serum proteomic analysis. One is a resolution challenge and results from the large number of different molecular species present in most biological samples. The second is a dynamic range challenge due to the considerable difference in concentrations between the most and least abundant molecular species present in the sample. (Finch 2008) As an example of this, consider a sample of human plasma: over 95% of the total mass of the proteins in plasma is made up of 12 species, namely human serum albumin, IgG, IgA, IgM, transferrin, fibrinogen, apolipoprotein A-I, apolipoprotein A-II, haptoglobin, α-1 antitrypsin, orosomucoid (α-1 acid glycoprotein), and α-2 macroglobulin. (Finch 2008) A breakdown of a small fraction of one of these peptides could overwhelm the detection of smaller, unknown peptides. (Skold 2002, Svensson 2003) Plasma and serum are highly complex body fluids that contain lipids, carbohydrates, amino acids, nucleic acids, hormones, and proteins. (Zubrig 2008 Biomarker) The information found in serum could describe the health state of an individual. (Thadikkaran 2005) The search for serum biomarkers is based on the assumption that "disease" correlates with an altered state of information circulating in the blood. (Zubrig 2008 Biomarker) Many types of molecules function as signaling molecules. It is likely that the low-molecular weight, low-abundance fraction of the human serum proteome is where the novel biomarkers are to be found. (Liotta 2003, Meng 2007)

It is hoped that the identity of potential biomarkers will provide insights into the pathology of the disease process. An NIH study group settled on a single definition of what a biomarker is: "A characteristic that is objectively measured and evaluated as indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention." (Biomarkers Group 2001) A biomarker is a molecule that allows for detection of changes in physiological state of an individual due to disease, treatment, or exposure to any other environmental factor that might affect the health of the individual. These insights could possibly lead to the development of new tools for detection and/or treatment. (Zubrig 2008 Biomarker) Biomarkers have the potential to identify pathological processes before they become symptomatic and could be used to screen for disease, even if that disease was at an asymptomatic stage. (Srinivas 2001) Biomarkers could also be used as prognostic markers to predict the progression of the disease, as stratification markers to predict the likely response to a treatment, or as efficacy markers to monitor the success or failure of treatment. (Zubrig 2008 Biomarker) When diseases have similar symptoms, as is the case with Alzheimer's disease and vascular dementia, biomarkers can be used to discriminate between them. (de Jong 2006) In most cases, proper diagnosis is required for effective treatment.

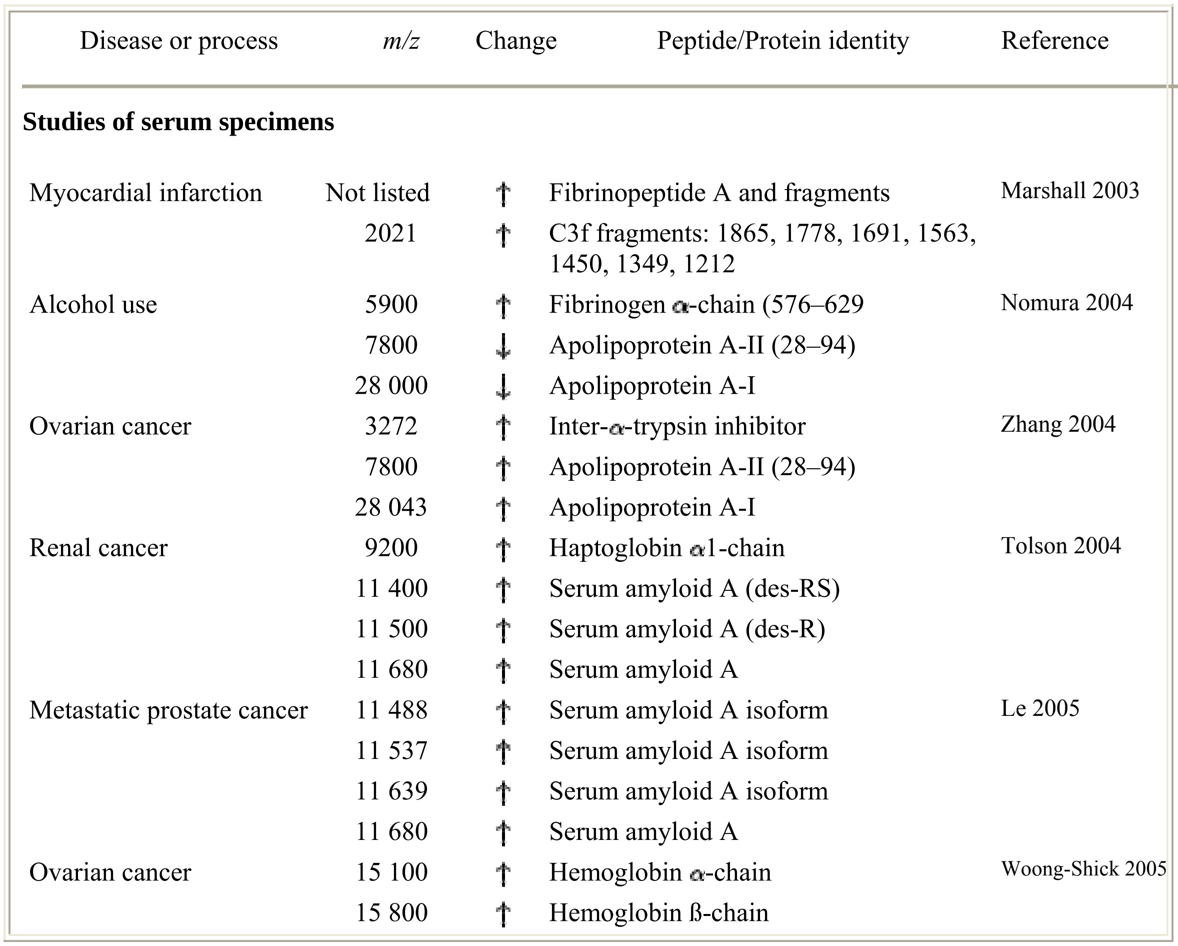
There are currently many biomarkers used diagnostically in medicine. One of the earliest and most widely known is prostate-specific antigen, still used as a screen for prostate cancer. There are an estimated 198,100 new cases of prostate cancer every year. This accounts for nearly 30% of all the cancers diagnosed in men. There were an estimated 39,200 deaths due to prostate cancer in 1998. (Grenelee 2001) Kuriyama et al developed an assay to measure prostate-specific antigen (PSA) in human serum and subsequent studies demonstrated the correlation between serum levels and prostate cancer incidence. (Kuriyama 1980, Catalona 1991) Since then, PSA has been the most commonly used tumor marker for prostate cancer. Conflicting results have been reported as to the accuracy of PSA as a biomarker with sensitivities (true positives) being reported in the 67-100% range with an average of 72% and reported specificities (true negatives) ranging from 18-100% with an average of 93%. The positive predictive value was reported in the 17-57% range with an average of 25%. These results suggest that when abnormal PSA findings are reported, there is a 1 in 4 chance that there is cancer, but when the PSA levels are normal the chance of missing cancer is about 1 in 10. (Kishor

2003) The currently accepted cutoff for PSA screening is a serum level of >4 ng/ml. Using this cutoff, 62% of prostate cancers are detected, but 38% of cancer patients have PSA below this level. This is partially due to increases in serum PSA levels due to non-cancerous prostate conditions, or advancing age. Regardless of the limitations, PSA serum screening has aided in early diagnosis for prostate cancer, and a decrease in patient mortality.

In the Surveillance, Epidemiology, and End Results (SEER) database, the percentage of prostate cancer patients with metastatic (distant) diseases at the time of diagnosis decreased from 16% in the period from 1985-1989 to 4% for the period from 1996-2004. The SEER data also show that the relative 5-year survival rate for metastatic prostate cancer is only 31.7%, while that for localized/regionalized prostate cancer is 98.9%. Indeed, the overall 5-year survival rate increased from 76.1% for the period 1984-1986 to 98.9% for the period 1996-2004. It is likely that the increased use of PSA-based screening is at least partially responsible for this marked improvement in both early diagnosis and overall survival rate. (Loeb 2007, Surveillance 2008) The continued success of PSA as a diagnostic screening tool despite its limitations has demonstrated the value of serum biomarkers for disease.

In rheumatoid arthritis (RA) it is important to discriminate between ‘slow’ and ‘fast’ progressing disease to allow treatment with appropriate drugs. If it is determined that a particular case of RA is of the faster progressing variety, more aggressive drugs are used to decrease the level of joint destruction in that patient. (Liao 2004) This assessment can be accomplished with biomarkers for RA. Biomarkers, in combination with the collagen turnover rate, can also be used to follow the effectiveness of RA treatment drugs over time. (Torikai 2006 Infliximab, Torikai 2006 Methotrexate) Early detection is crucial for control and prevention of many diseases. These diseases include, but are not limited to RA, Parkinson's disease, Alzheimer's disease, and cancer. (Dorsey 2006, Sunderland 2006, Wulfkuhle 2003)

Because the potential for biomarkers is so great, many groups are searching for biomarkers for many different diseases. A compilation of the serum/plasma biomarkers identified in 24 studies using MS is shown in Table 1.1. Many of the biomarkers are fragments or isoforms of larger and abundant serum proteins. This extensive, but not comprehensive, list demonstrates the great strides that have been made in recent years towards novel biomarker discovery.



Disease or process	*m/z*	Change	Peptide/Protein identity	Reference
Studies of serum specimens				
Myocardial infarction	Not listed	↑	Fibrinopeptide A and fragments	Marshall 2003
	2021	↑	C3f fragments: 1865, 1778, 1691, 1563, 1450, 1349, 1212	
Alcohol use	5900	↑	Fibrinogen α-chain (576–629	Nomura 2004
	7800	↓	Apolipoprotein A-II (28–94)	
	28 000	↓	Apolipoprotein A-I	
Ovarian cancer	3272	↑	Inter-α-trypsin inhibitor	Zhang 2004
	7800	↑	Apolipoprotein A-II (28–94)	
	28 043	↑	Apolipoprotein A-I	
Renal cancer	9200	↑	Haptoglobin α1-chain	Tolson 2004
	11 400	↑	Serum amyloid A (des-RS)	
	11 500	↑	Serum amyloid A (des-R)	
	11 680	↑	Serum amyloid A	
Metastatic prostate cancer	11 488	↑	Serum amyloid A isoform	Le 2005
	11 537	↑	Serum amyloid A isoform	
	11 639	↑	Serum amyloid A isoform	
	11 680	↑	Serum amyloid A	
Ovarian cancer	15 100	↑	Hemoglobin α-chain	Woong-Shick 2005
	15 800	↑	Hemoglobin ß-chain	

Prostatic disorders	8946	↑	Apolipoprotein A-II	Malik 2005
Nasopharyngeal cancer	11 600	↑	Serum amyloid A	Cho 2004
	11 800	↑	Serum amyloid A	
Breast cancer	8116	↑	C3a (des-9 residues)	Li 2005
	8926	↑	C3a (des-R)	
Retrovirus infection	11 700	↑	α_1-Antitrypsin fragment	Semmes 2005
	11 950	↑	Haptoglobin-2 fragment	
	19 872	↑	Haptoglobin-2 fragment	
Breast cancer	6647	↓	Apolipoprotein C-I	Goncalves 2006
	8936	↓	C3a	
	9192	↑	Haptoglobin α_1-chain	
	28 284	↓	Apolipoprotein A-1	
	81 763	↑	Transferrin	
Diet change	2740	↑	α_2-HS-glycoprotein B-chain	Mitchell 2005
Benzene exposure	~7700	↓	Platelet factor-4	Vermeulen 2005
	~9300	↓	Connective tissue-activating peptide-III	
Ovarian cancer	13 900	↓	Transthyretin and ~12 900	Kozak 2005
	15 900	↑	Hemoglobin ß-chain	
	28 000	↓	Apolipoprotein A-I	
	79 000	↓	Transferrin	
Cancers (prostate, bladder, or breast)	1537	↓	Fibrinopeptide A and fragments: 1466, 1351, 1264, 1207, 1078, 1020, 906, 758	Villanueva 2006
	3261	↓↑	Fibrinogen α-chain fragments: 3190, 2931, 2768, 2553, 2379, 2816, 3239, 2659	
	2021	↑↓	C3 fragments: 1865, 1778, 1751, 1691, 1563, 1450, 1349, 1212, 1056, 942	
	3201	↑↓	C4a fragments: 2704, 2305, 1763	
	1896	↑↓	C4a fragments: 1740, 1627, 1499	
	3971	↑	Inter-α-trypsin inhibitor fragments: 3272, 2724, 2627, 2358, 2271, 2184, 2028, 1789, 998, 842 (or ↓)	
	3157	↑	Inter-α-trypsin inhibitor fragments: 2115	
	3377	None	Apolipoprotein A-I fragments: 1808	
	3182	↑	Apolipoprotein A-I fragment and 1971	

Table 1.1 Biomarkers identified in 24 published studies using MALDI-TOF or SELDI-TOF MS (18 studies examined serum specimens and 6 studies examined plasma). (Hortin 2006)

In order to preserve the integrity of the information in a serum sample, the fewest number of processing steps possible should be employed. In fact, the ideal situation would be one in which a crude, unprocessed sample is used for investigative analysis. Unfortunately, this is not possible with most body fluids. Easily accessible body fluids like serum have proteins that are already in solution, but serum also contains a complex mixture of molecules covering a wide range in size, concentration, polarity, and hydrophobicity. Because of the extreme range in concentration of the constituent molecules found in serum, the most abundant molecules must be removed to be able to visualize the less abundant molecules using MS. Removal of the larger highly abundant proteins, such as albumin, can be accomplished using affinity columns (Zolotarjova 2005, Bailes 2008), ultrafiltration, or organic solvent precipitation (Merrell 2004). Other methods used to process samples include size exclusion, anion exchange (Kaiser 2003), and reverse-phase separation (Wittke 2003). After sample preparation, a separation step is generally employed to reduce the complexity of the sample to maximize the sensitivity of ionization and subsequent MS analysis. (Zubrig 2008 Peptidomics)

Peptidomics

Peptides, like proteins, are formed of chains of various amino acids linked together through amide bonds. The main difference between proteins and peptides is the number of amino acids in the chain; proteins have more and are long, and peptides have fewer and so are short. There is no specific, accepted length at which a polymer of amino acids ceases to be a peptide and begins to be a protein. Some people apply a cutoff length of approximately 50 amino acids in length, but this has been disputed. Some ‘proteins’ (such as insulin) fall close to the upper limit of this definition, and some peptides (e.g. amyloid beta peptide) are often considered ‘proteins’ despite falling below this cutoff. Another suggested way to differentiate between proteins is by mass; with proteins having masses above 10 kDa and peptides having masses below 5 kDa. (Schrader 2001, Schulz-Knappe 2005) However, this definition is not universally accepted or applicable; ubiquitin is considered a protein and has a mass of just 8 kDa, many peptide hormones like insulin and IGF have masses below 10kDa, others like PDGF have larger masses. (Cutillas 2008) Because both of these considerations are somewhat arbitrary, it has been suggested that the scientific community as a whole ascribe to the definition that "a peptide is a poly-amino acid molecule without tertiary structure; on gaining defined structure, it is a protein." (Zubrig 2008 Peptidomics) This has yet to become the standard.

The first successful studies of the peptidomes of various biological samples were reported over 60 years ago. These samples included urine (Baar 1956), blood (Jolley 1965), and brain tissue (Hughes 1975) and were accomplished using mainly chromatography and hard-ionization pre-LC-MS techniques. (Lucas 1969) The study of peptides (peptidomics) is important in the search for novel biomarkers in serum. Specific kinds of peptides function in cell to cell communication in an autocrine, paracrine, or endocrine (systemic) manner. (Strand 2003) In addition, since peptides can often redistribute between the various compartments of the body, it is probable that peptide content within the serum will mirror changes made by disease. (Zubrig 2008 Peptidomics) The serum peptidome then provides an integrated sample, being comprised of secreted and intracellular peptides from throughout the body. (Griffiths 2008) It is likely that the low-abundance fraction of the human serum peptidome will prove useful in the search for novel biomarkers for disease. (Liotta 2003, Meng 2007)

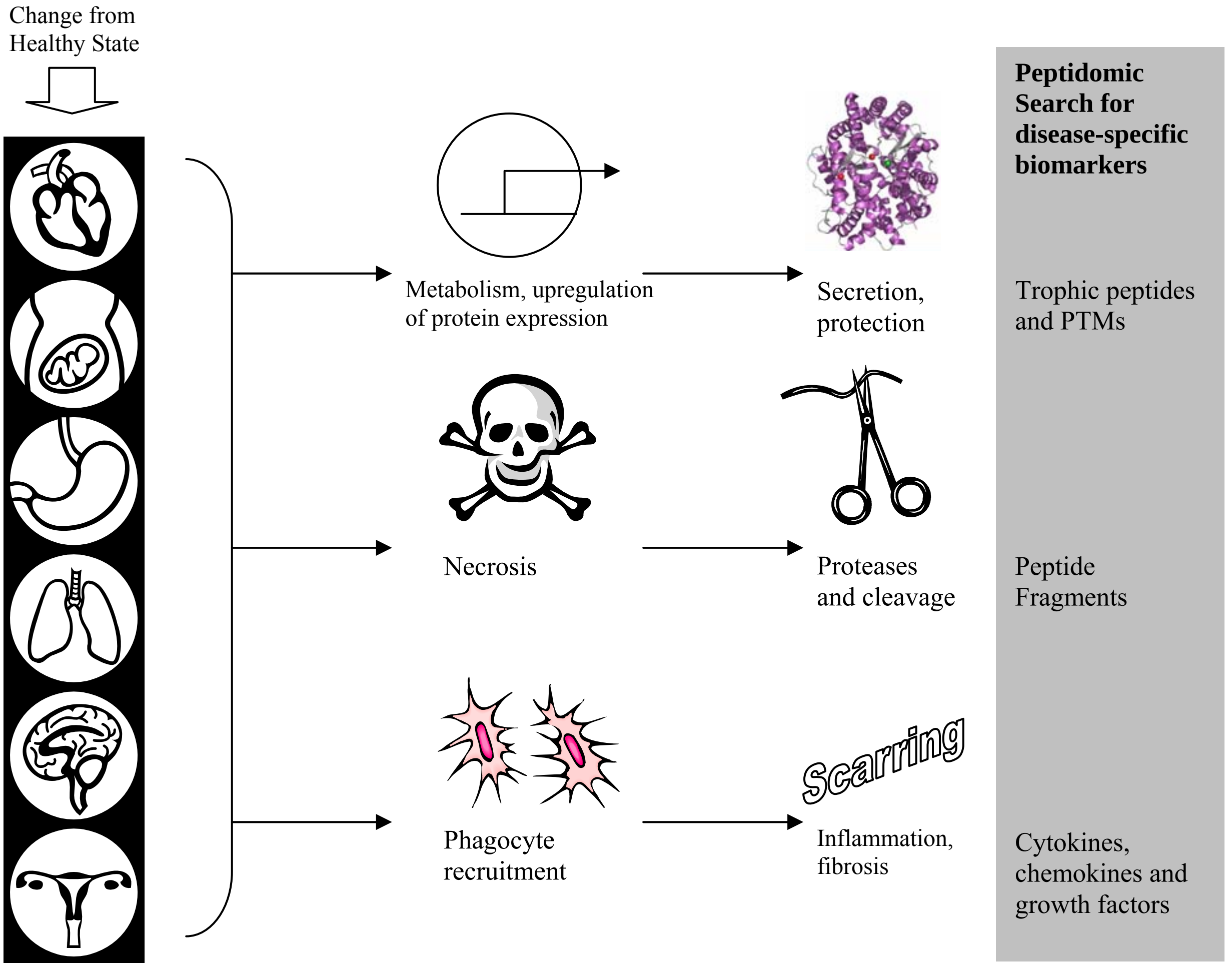


Figure 1.3. The possible sources of disease-specific peptides in the serum proteome. Disease causes a change in tissues and/or organs from their normal healthy state. Disease may induce posttranslational modifications (PTMs), proteolytic cleavage, or other responses. (Adapted from Griffiths 2008)

Thousands of peptides have been tentatively identified in plasma. (Hortin 2006, Omenn 2005) The majority of the peptides in serum samples are bound to carrier molecules, primarily serum albumin. (Araujo 2008) The abundance of carrier molecules is high enough compared with the concentration of free peptides in the serum that even if the binding affinity between a carrier molecule and a free peptide is low, the peptide is more likely to exist in a bound form than a free form in the serum. (Figure 1.4) Because larger proteins have longer half-lives in the serum (Table 1.2), the binding of smaller peptides to

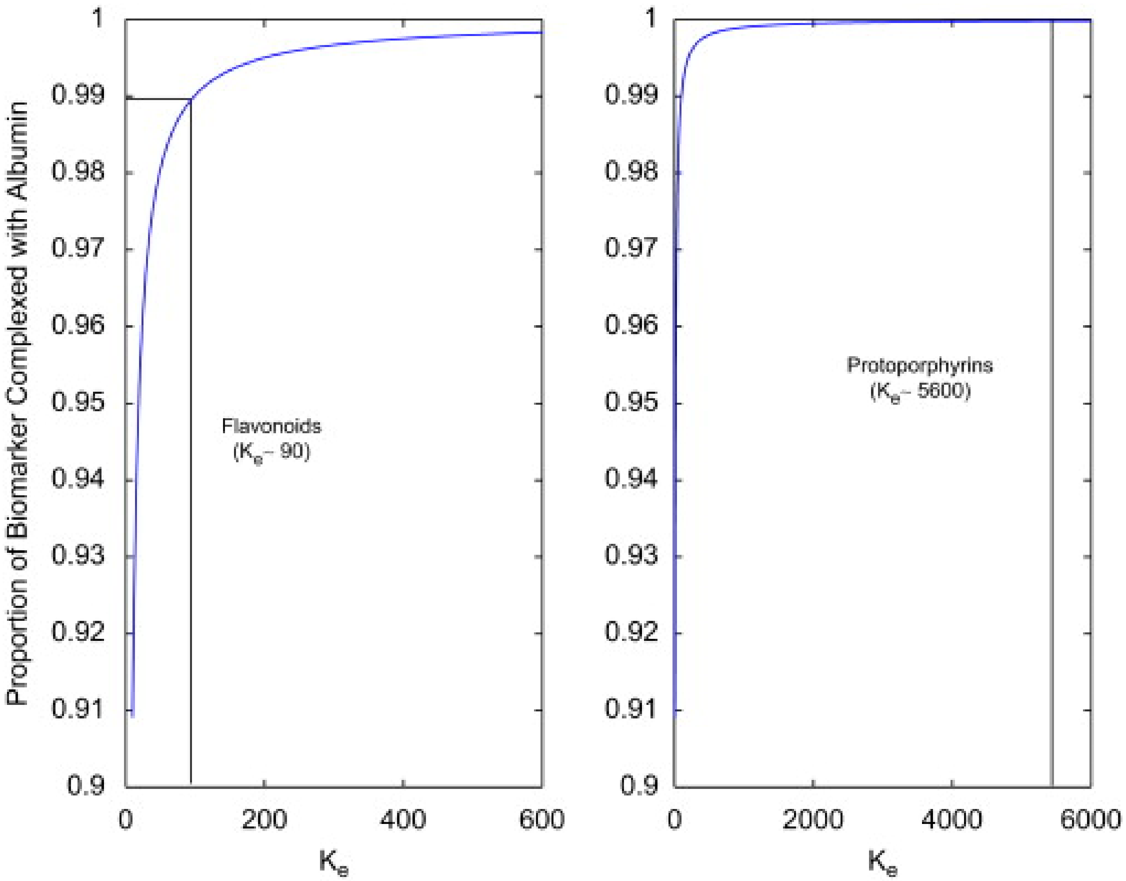


Figure. 1.4 Low-abundance molecules such as flavonoids (left) and protoporphyrins (right), with low-affinity for albumin, will nevertheless exist almost entirely in complexed association with albumin in the bloodstream on account of albumin's overwhelming abundance. The large albumin concentration amplifies the *effective* affinity constant, K_e, to K_e∼90 for flavonoids and K_e∼5600 for protoporphyrins, corresponding to 98.9% and 99.98% probability that the biomarker will be complexed with albumin for flavonoids and protoporphyrins, respectively. (Araujo 2008)

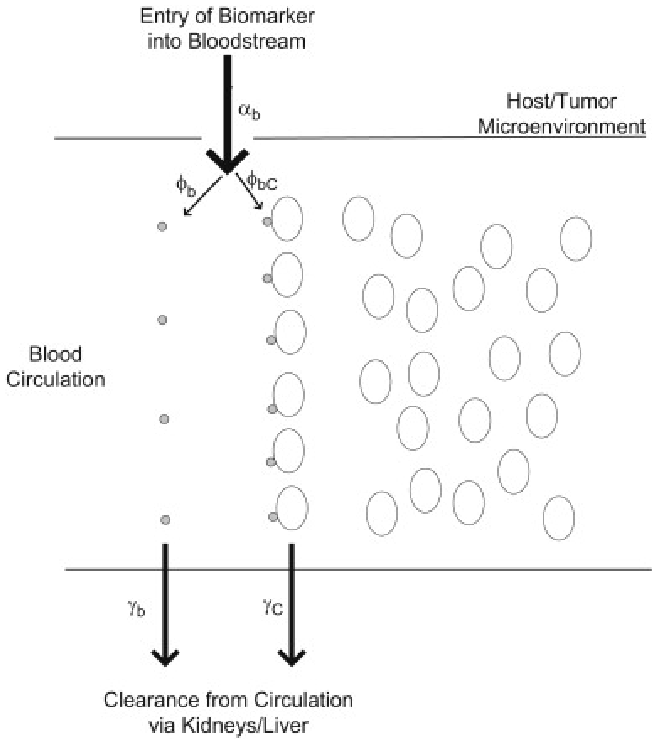
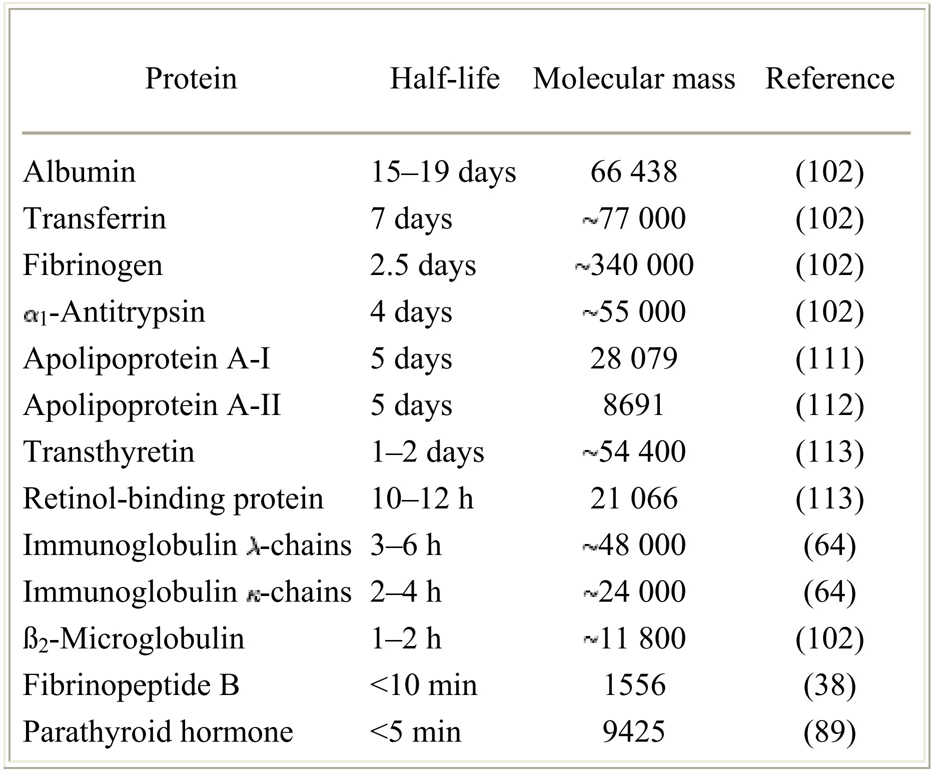

Entry of Biomarker into Bloodstream
α_b
Host/Tumor Microenvironment
ϕ_b
ϕ_{bC}
Blood Circulation
γ_b
γ_C
Clearance from Circulation via Kidneys/Liver

Fig. 1.5. Biomarkers in the blood stream. A low-abundance biomarker enters the blood stream at a constant rate, α_b. Once in the bloodstream, a proportion of the biomarker will bind to carrier proteins (φ_{bC}, while the remainder resides in the bloodstream in a free (unbound) form (φ_b). (Araujo 2008)

larger ones gradually but effectively amplifies their total concentration in serum. Peptides that do not bind to larger molecules are cleared from the serum by the kidneys within hours, or even minutes. (Figure 1.5) Changes in disease state that are reflected in the serum will not be instantly reflected in the total levels of the biomarker concentration. Rather, the secretion of biomarker into the serum by diseased tissue will gradually be bound by carrier molecules, increasing their half-life dramatically and slowly increasing their total concentration. (Araujo 2008) These complex mixtures of bound peptides have not been thoroughly studied, but if each molecule of albumin binds 1 peptide, there exists a potential capacity of more than 500 μmol/L of bound peptides. (Omenn 2005) Thus, when searching the serum proteome for potential biomarkers, it is essential to include that fraction that is bound to larger carrier proteins.



Protein	Half-life	Molecular mass	Reference
Albumin	15–19 days	66 438	(102)
Transferrin	7 days	~77 000	(102)
Fibrinogen	2.5 days	~340 000	(102)
α_1-Antitrypsin	4 days	~55 000	(102)
Apolipoprotein A-I	5 days	28 079	(111)
Apolipoprotein A-II	5 days	8691	(112)
Transthyretin	1–2 days	~54 400	(113)
Retinol-binding protein	10–12 h	21 066	(113)
Immunoglobulin λ-chains	3–6 h	~48 000	(64)
Immunoglobulin κ-chains	2–4 h	~24 000	(64)
ß$_2$-Microglobulin	1–2 h	~11 800	(102)
Fibrinopeptide B	<10 min	1556	(38)
Parathyroid hormone	<5 min	9425	(89)

Table 1.2. Half-lives of proteins and peptides in the circulation. (Richter 1999)

Summary

This introduction has outlined both the ever-increasing need for novel biomarkers for disease and some of the technologies currently employed to identify prospective biomarkers. The large number of biomarkers reported in Table 1 show that previously described methods have produced a promising number of potential candidates, but not necessarily a high number of exciting results. Most of the potential biomarkers identified are fragments of large, highly abundant proteins found within the blood. Many of these are well-characterized and not likely to provide new insights into the pathology, progression, or treatment of disease. Novel biomarkers that are less abundant are likely to be more informative. The low molecular weight component of the serum proteome is a likely place to begin the search for novel and informative biomarkers for disease. For this reason, we have chosen to focus our search for novel biomarkers within this subset.

Several of the challenges associated with the search for biomarkers in the low-abundance, low-molecular weight fraction of the serum will be discussed in detail in the following chapters. The second chapter outlines several methods that were developed to address issues encountered in sample preparation, data normalization and alignment, and identification of potential biomarker peptides using MSMS. The third and fourth chapters are examples of our search for biomarkers for preterm birth and preeclampsia. While our initial results seem promising, we realize that what has been found so far is just the tip of the iceberg when it comes to potential biomarkers for these conditions. There is still much to be done with regards to searching, sequencing, and validating these potential markers in order for them to be useful in a diagnostic setting, but these results represent a very good start towards a non-invasive test for these and other diseases that are very hard to diagnose very far in advance of onset of the diseases.

Chapter 2 – Methods

Part I – Analysis of Low-Abundance, Low-Molecular-Weight Serum Proteins Using Mass Spectrometry

Abstract

To detect diseases early in the general population, new diagnostic approaches are needed that have adequate sensitivity and specificity. Recent studies have used mass spectrometry (MS) to identify a serum proteomic pattern for breast and ovarian cancer. Serum contains 60-80 mg protein/mL, but 57-71% of this is serum albumin, and 8-26% is γ-globulins. These large proteins must be depleted before smaller less-abundant proteins can be detected using MS, but because serum albumin is known to act as a carrier for smaller proteins, removal of these molecules using columns or filtration may result in the loss of molecules of interest. The objective of this study was to develop a reproducible method to deplete serum samples of high-abundance proteins in order to analyze the less-abundant proteins present in serum. We used organic solvents to precipitate the large proteins out of solution. We also predicted that this would cause many smaller proteins to dissociate from their carrier molecules, allowing for detection of a larger number of small peptide and proteins. These treated samples were analyzed using capillary liquid chromatography coupled with electrospray ionization mass spectrometry. Analysis demonstrated reproducible results. Acetonitrile treatment clearly released many carrier bound molecular species and was superior to ultrafiltration alone for serum proteomic analysis.

Introduction

Many diseases, such as cancer, do not show obvious clinical symptoms until the disease is in advanced stages and difficult to treat successfully. Detection of such diseases in early stages can greatly reduce disease-related mortality. (Li 2002, Petricoin 2003 Clinical) Traditionally, researchers have searched for disease biomarkers using a one-at-a-time approach. However, because of the complexity of many diseases, it is very likely that the use of multiple biomarkers (protein pattern diagnostics) in screening and diagnosis will be necessary to produce unequivocal results. (Li 2002, Petricoin 2002 Clinical)

Biomarkers are often low molecular weight proteins secreted into the blood stream as a result of the disease process. (Petricoin 2002 Clinical) Employing mass spectrometry (MS) in serum proteomic pattern diagnostics has a very high potential for discovering new, relevant disease-related biomarkers. (Petricoin 2003 Counterpoint) Serum contains a total of 60–80 mg/mL of protein, with a concentration range spanning at least nine orders of magnitude. An estimated 10,000 different proteins typically exist in serum, the majority of which are low molecular weight species. (Adkins 2002) The complexity of serum makes it a very informative source for developing a proteomic pattern, however, 65–97% of serum protein is serum albumin and immunoglobulin. The presence of these highly abundant proteins makes detection of the smaller, less-abundant serum proteins difficult. (Adkins 2002, Govorukhina 2003)

Some have used cibacron dye and protein A/G columns to remove serum albumin and immunoglobulin from serum effectively, allowing detection of biomarkers present at very low concentrations. (Govorukhina 2003, Ahmed 2003) Molecular weight cutoff filters have also been evaluated in separating the large, abundant proteins from the smaller, less abundant ones. (Georgiou 2001) Albumin serves as a transport protein that binds, and thereby accumulates many low-abundance biomarkers that would otherwise be cleared from the blood by the kidney. (Petricoin 2002, Adkins 2002, Liotta 2003) Consequently, removal of serum albumin by filtration also inadvertently removes many small proteins and peptides of interest.

Matrix-assisted laser desorption ionization mass spectrometry (MALDI-MS), surface-enhanced laser desorption ionization mass spectrometry (SELDI-MS), and liquid chromatography-coupled to online electrospray mass spectrometry (cLC-MS) are methods that have been used for protein pattern development for several diseases. MALDI-MS employs no separation, and it is sensitive to organic and inorganic contaminants as evidenced by signal background problems. (Li 2002) SELDI-MS is an affinity-based method that uses various stationary sorbants (hydrophobic, ionic, cationic, and metal binding) to selectively bind proteins of interest. Because many biomarkers are present at very low concentrations in biological fluids, they are not likely to be immobilized on the biochips used in SELDI-MS but may be washed away with all of the other protein species that do not bind to the active surface. (Diamandis 2003, Banez 2003) Current SELDI instruments are also severely limited in their resolution. cLC-MS, while not as high throughput as MALDI-MS or SELDI-MS, is more effective for the detection of low-molecular weight, low-abundance proteins and peptides in a complex mixture.

In this report, we hypothesized that using acetonitrile precipitation, which has been shown to effectively precipitate large, abundant proteins out of serum, (Alpert 2003) is a superior method to remove serum albumin and immunoglobulin from serum because acetonitrile not only denatures these large proteins allowing for their removal but should also cause the small, low-abundance proteins and peptides normally bound to these carrier proteins to dissociate, thus making them available for detection using cLC-MS. We report that this procedure was effective and allowed for the visualization of small proteins and peptides from serum. An average of ~4,000 ionized species under 2.5 kDa, as well as some proteins between 2.5 and 10 kDa, were detected per serum sample analyzed.

Materials and Methods

All reagents were purchased from Sigma (Saint Louis, MO)

Sample Collection

After Institutional Review Board (IRB) approval was obtained at the University of Utah (IRB # 10535), serum samples were obtained from women who presented to the labor and delivery suite with a pregnancy at ≥ 36 weeks gestation. Blood was obtained by antecubital venipuncture and after centrifugation at 3,500 rpm for 15 minutes, serum was collected, aliquoted and frozen at -80º C until further processing. Samples were also obtained at the BYU Student Health Center from healthy male volunteers.

Acetonitrile Precipitation

Briefly, the ‘standard’ method (Alpert 2003) involved adding two volumes of HPLC grade acetonitrile to 1 volume serum followed by gentle mixing. After standing at room temperature for 30 minutes, samples were spun for 4 minutes at 12K rpm. An aliquot of supernatant was then lyophilized to 50 µl. The sample was reconstituted to original volume by addition of water.

The ‘modified’ method involved adding two volumes of HPLC grade acetonitrile (400 µl) were added to 200 µl serum, vortexed vigorously for 5 seconds, and allowed to stand at room temperature for 30 minutes. Samples were then spun for 10 minutes at 12K rpm in an IEC Micromax RF centrifuge at room temperature. An aliquot of supernatant (550 µl) was then lyophilized to 200 µl in a vacuum centrifuge (LABCONCO CentriVap Concentrator). Supernatant protein concentration was determined using a Bio-Rad microtiter plate protein assay according to manufacturer instructions. An aliquot containing 4 µg protein was transferred to a new microcentrifuge tube and lyophilized to near dryness. Then 20 µl 88% formic acid, 2 µl 5 pmol/µl mellitin, and 2 µl 5 pmol/µl [Glu^1] fibrinopeptide were added to lyophilized samples. Samples were brought to 40 µl with HPLC water.

Centrifugal Ultrafiltration

1 ml of serum was diluted with 5 ml of 25 mM ammonium bicarbonate and added to an appropriately conditioned 30 K molecular weight cutoff (NMWL) centrifugal filter. After centrifugation, the filtrate was lyophilized to dryness and reconstituted to 40 µl with 20 µl 88% formic acid and 20 µl water. The retentate was reconstituted to its original 1 ml volume with HPLC grade water. Additionally, 200 µl of retentate were subjected to acetonitrile precipitation and 1 µg of protein of each sample was loaded onto the column according to procedures outlined below.

cLC-MS Analysis

Capillary liquid chromatography, to fractionate or separate peptides and proteins, was performed using a 15 cm x 250 um i.d. capillary column, packed in-house using POROS R1 reversed-phase media (Applied Biosystems, Framingham MA), employing a

2.2%/min gradient to an organic concentration of 60% acetonitrile in 0.1% formic acid, followed by a 3.5%/min gradient up to a concentration of 95% organic phase. Chromatography used an LC Packings Ultimate Capillary HPLC pump system, with a FamOS autosampler (Dionex Corp), controlled by the mass spectrometer software (Analyst). The LC is coupled directly to the MS. Effluent from the capillary column was directed into a QSTAR Pulsar i quadrupole orthogonal time-of-flight mass spectrometer through an IonSpray source (Applied Biosystems). Data were collected for m/z 500 to 2500 over the entire range of the chromatogram (55 min). Data collection, processing and preliminary formatting was accomplished using the Analyst QS software package with BioAnalyst add-ons (Applied Biosystems).

Normalization of Elution Times

Small differences in elution time were accounted for by normalizing them to the elution time of a molecular species with a unique m/z value, common to all samples.

Results and Discussion

Evaluation of Protein Precipitation Approaches Our goal was to develop a protein precipitation method that had sufficient reproducibility to allow for accurate quantitative serum proteomic analysis in an effort to identify molecular species that differed between healthy and diseased states. The adequacy of this method was tested using cLC-MS. Additionally, we were interested in determining the residual molecular species and their molecular weights after a precipitation procedure.

Organic solvents have long been used to precipitate proteins. Unlike other techniques, such as ultrafiltration or dialysis, this technique will typically result in many smaller proteins and peptides being released from large carrier proteins.

In our initial study two techniques were evaluated. The first was a previously described method that represented one of the first approaches to serum preparation for global serum proteomic analysis using MS. The second was a modification of that technique. The methods are described in detail in the Methods section and summarized in Table 2.1. The results of the two approaches are shown in Figure 2.1. The previously described or 'standard' method resulted in a mean protein level of 0.240 ± 0.124 mg/ml. The modified method produced a mean protein concentration of 0.028 ± 0.0006 mg/ml which was significantly less than the 'standard' method (Students Unpaired t-test: $P=3.1 \times 10^{-5}$). Moreover, protein concentrations remaining after the modified method were far less variable (CV method 1 = 51.7%, CV method 2 = 21.4%). This reduction was an improvement, but admittedly did not completely eliminate variability. It is beyond the

Standard Method	**Modified Method**
Mixing after addition of Acetonitrile	
Gentle Mixing	Vortexing
cloudy supernatant	clearer supernatant
higher protein concentration	lower protein concentration
inconsistent protein concentration	consistent protein concentration
Centifugation	
4 min. at 12,000 rpm	10 min. at 12,000 rpm
cloudy supernatant	clearer supernatant
Concentrated Volume of Supernatant	
50 µl	200 µl
precipitate forms	no precipitate forms

Table 2.1. Optimization of acetonitrile precipitation. The table outlines the two methods tested and their respective relevant findings.

aims and scope of these experiments to determine whether the two methods resulted in different proteins being precipitated or in more efficient removal of the same proteins.

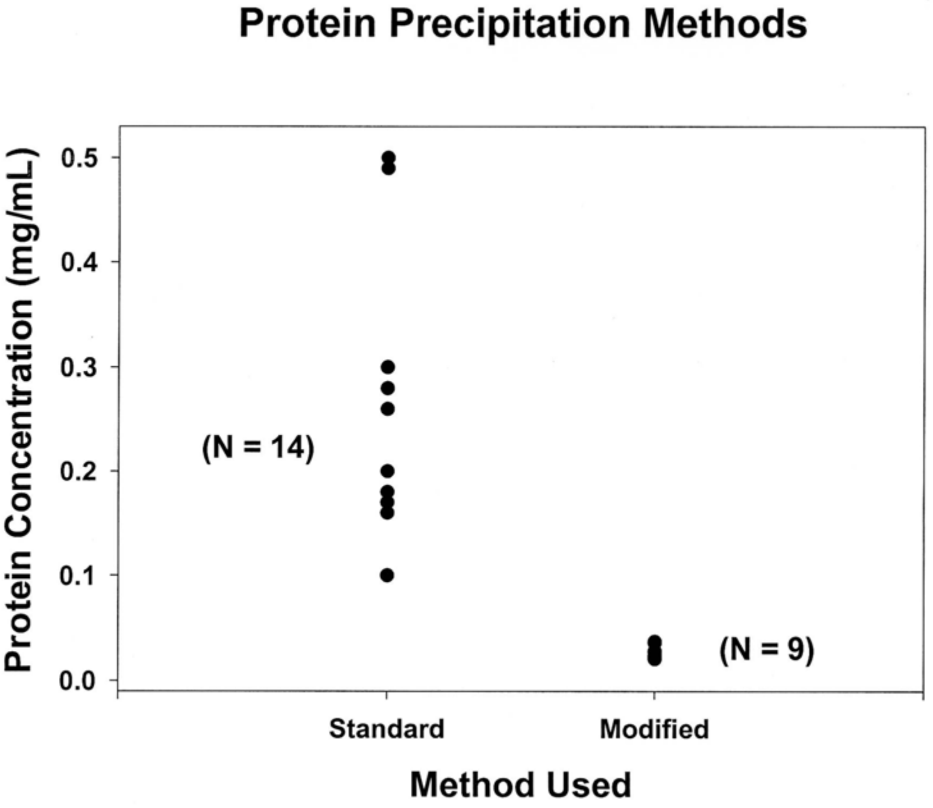


Figure 2.1. Protein yield comparison of standard and modified acetonitrile precipitation. The left column shows that the standard method varied greatly in protein yield. As shown in the column on the right, the new modified method resulted in a more reproducible protein yield.

In addition, the original method left a cloudy solution after centrifugation, suggesting that some of the variability may result from an inadequately long or forceful centrifugation step. This unresolved precipitate is likely to be difficult to measure accurately in a protein assay and may contribute to the marked variability of residual protein concentration.

Further reductions in the liquid portion of the supernatant, in particular taking the sample to dryness, resulted in a residue that was no longer appreciably soluble in the cLC mobile phase solvent.

cLC-MS Evaluation of Residual Proteins

Because of the improved reproducibility of the modified method, it was used for all subsequent experiments. The final volume of 200 µl was too large to inject onto our capillary column. Therefore, immediately prior to specimen introduction into the cLC-MS system, a volume of solution containing 4.0 µg of protein was aliquotted and concentrated to near dryness, avoiding complete dryness. Any precipitate that formed upon concentration was resolubilized by the addition of 20 µl of 88% formic acid. There was no appreciable protein loss in this step (data not shown).

This precipitation method was further evaluated to determine the molecular weight range and number of molecular species present after treatment. Given the large number of molecular species, the whole elution interval was divided into five minute time frames and the mass spectra were averaged for that period. Then the LCMS reconstruct tool in BioAnalyst was applied to the averaged spectrum. Output from the several 5 minute intervals was compiled to capture all species over the entire sample run. Figure 2.2 represents the average of 49 different specimens. While the most abundant species were below a MW of 2,000 (Panel A), many species with MW >2,000 daltons (Panel B) were still present.

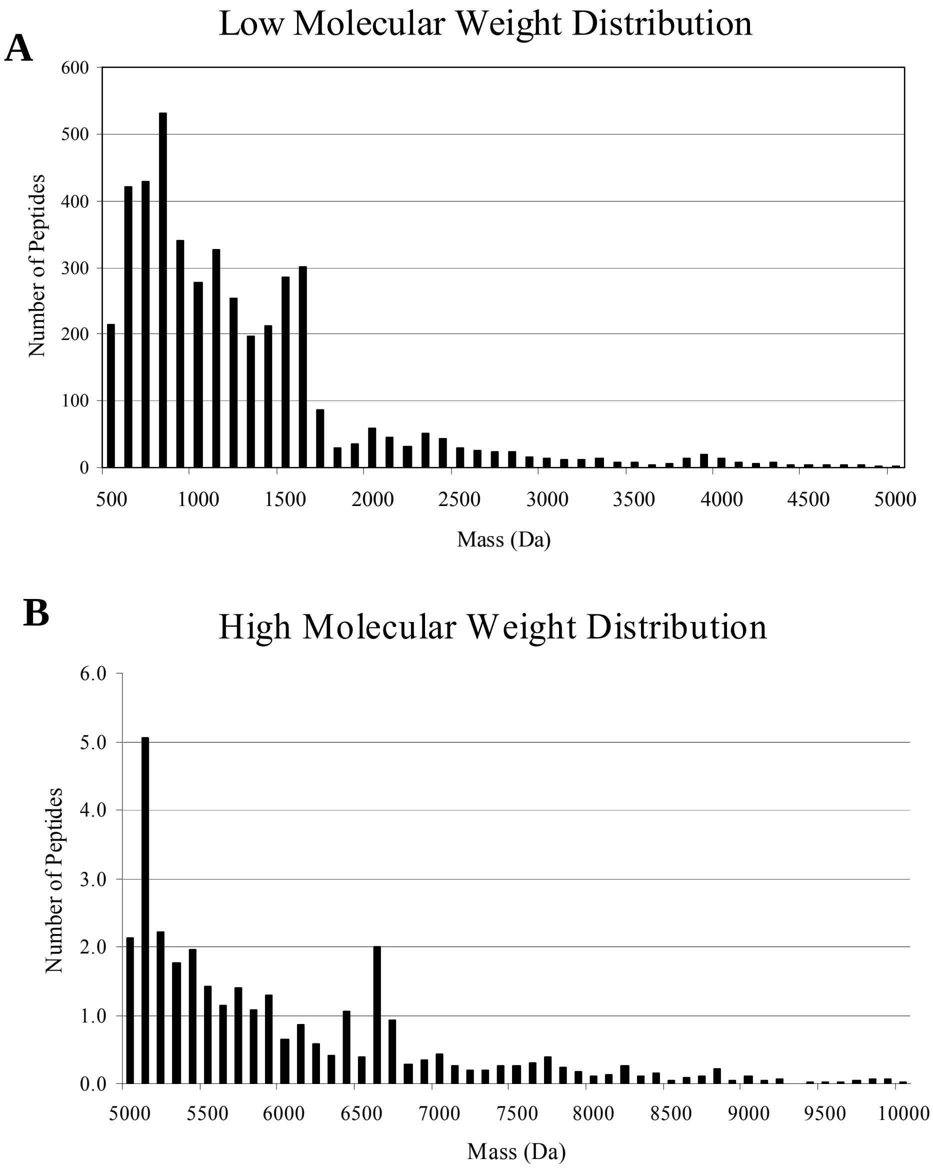


Figure 2.2. **Molecular weight distribution of species observed in acetonitrile supernatant.** cLC-MS was used to analyze 49 serum specimens after acetonitrile precipitation. A) This histogram shows the distribution of all molecular species ranging from 500 to 5000 Daltons. B) This histogram shows the distribution of all molecular species ranging from 5 to 10 kD. The majority of the species present were below 2000 Daltons.

It is important to note that close comparison of the actual unmanipulated mass spectra to the software LCMS reconstructed data demonstrated that the LCMS reconstruct tool occasionally omitted species, and did so variably, making it unreliable for use in clinical studies where quantification is critical.

Value of ACN Precipitation

The effects of acetonitrile precipitation of serum proteins were evaluated by direct comparison of pre- and post-precipitation cLC-MS chromatograms. Comparisons of untreated and treated serum demonstrated that many more molecular species were observable by MS after treatment. This is evident in panels A and B of Figure 2.3 which show an expanded number of molecular species in the total-ion chromatogram (TIC) after treatment (panel B). Moreover, the mass spectrum of the untreated specimen was dominated by the extensive envelope (See panel C), which peaked between an m/z of 1300-1400, representing a species with molecular weight of 60-70 kDa (most likely albumin or related compounds) which masked many molecular species. Even the region of lower m/z (as shown in the insert of panel D) had more species revealed by ACN treatment. These findings appear to confirm the hypothesis that organic solvent precipitation of proteins would also release smaller, protein-bound species. While it is true that larger proteins present in the serum specimen are sacrificed by ACN treatment, it is also true (as clearly evident in panel E) that higher molecular weight species are obscured by the highly abundant albumin and that lower molecular weight species are both more abundant and more observable after treatment (see panel C and D as well as E). In summary, the ACN pre-treatment actually provides broader mass spectral access to the serum proteome and makes more likely the development of an informative and diagnostic pattern for defining differing states in the human. It should also be noted that most growth, stimulatory, regulatory and pathology-responsive factors are small proteins and peptides. Consequently, the additional ability of serum proteomics to actually define the mediators of disease may be more successful after treatments that more fully expose this lower molecular weight sub-proteome.

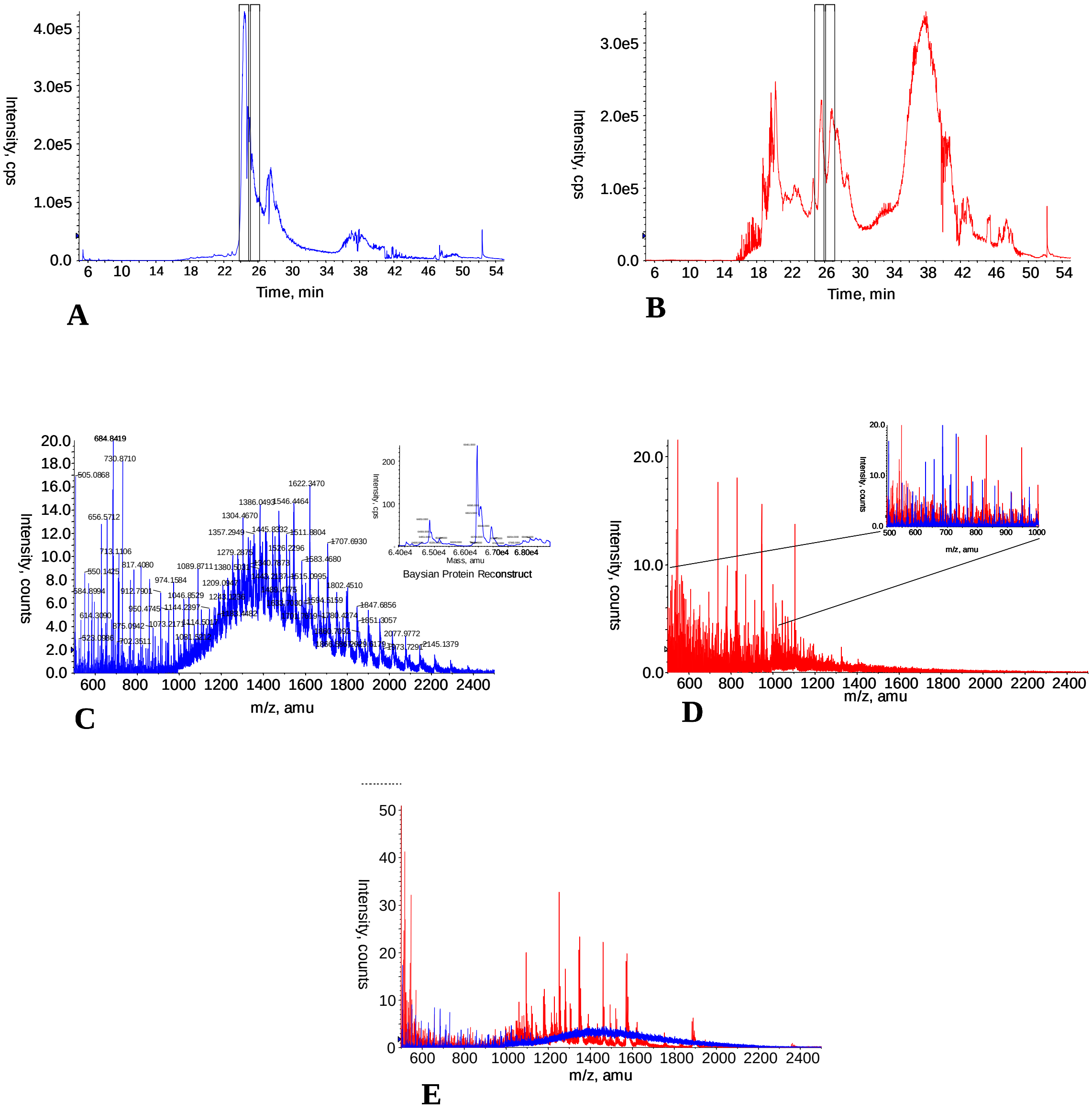


Figure 2.3. **Comparison of serum protein (1µg) before and after acetonitrile precipitation.** Total ion chromatograms (TIC) for serum proteins before (A) and after (B) acetonitrile are presented. The averaged mass spectra for the cLC elution (taken from 24 to 25 min) of a serum specimen prior to ACN precipitation (C) and after ACN precipitation (D) are shown. Panel C contains an inlay that shows the protein mass graph of the Baysein protein reconstruct performed on the mass spectrum from 900 m/z to 2500 m/z. The inlay in panel D shows the expanded (500 to 1000 m/z) mass spectra overlay of the pre and post samples. E) The overlay of the mass spectra for species present between 25 and 26 min of elution for the two samples is shown. Before ACN precipitation is shown in blue, after ACN precipitation is shown in red.

Ultrafiltration vs. Precipitation

Centrifugal ultrafiltration is a common method used to remove serum albumin, immunoglobulins and other abundant, high molecular weight proteins from a serum specimen. (Tirumalai 2003) Such treatment is often necessary before chromatography or capillary electrophoretic techniques can be employed. Although this method is selective for low molecular weight peptides and proteins in a sample, the resulting filtrate may not allow for a thorough examination of the entire low molecular weight serum proteome. Small proteins and peptides are commonly bound by large carrier proteins to both moderate their free and biologically active concentration and to prevent them being rapidly cleared from the blood by the kidney. The experiments described in the previous section suggested strongly that there was an increase in low molecular weight species

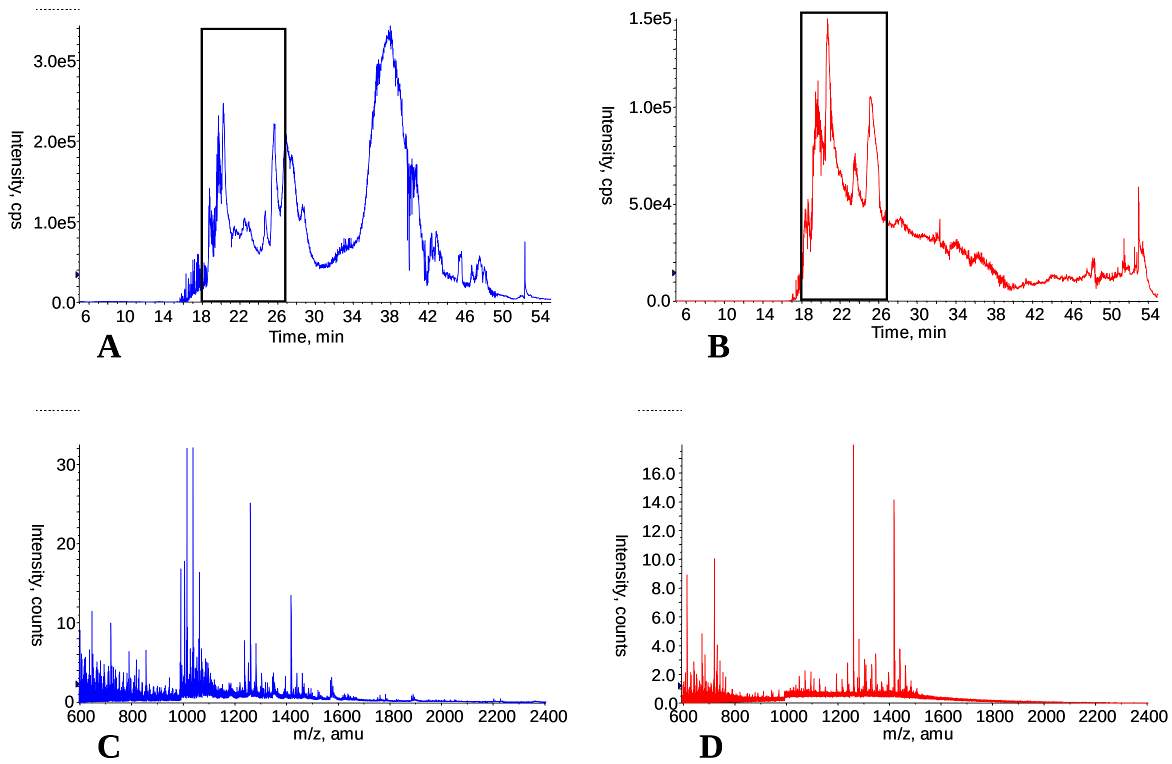


Figure 2.4. LCMS of acetonitrile supernatant compared to 30K NMW filtrate. Total ion chromatograms of ACN-precipitated serum (A) and reconstituted filtrate (B) are shown. Panels C and D show the averaged mass spectra overlay from 18 to 27 min cLC elution for the serum and the filtrate respectively. ACN-precipitated serum is shown in blue, the filtrate is shown in red. Comparison of panels C and D, shows that the ACN precipitation method was superior to the centrifugal ultrafiltration technique in both number of low molecular weight species and intensity of peaks.

after the addition of ACN to denature serum proteins. A second line of evidence for this is provided by comparison of the ACN precipitation with ultrafiltration of the same serum.

Figure 2.4 provides cLC-MS results for equal quantity protein injections (1 μg) of post ACN-precipitated serum (panel A) with the filtrate obtained after a 30,000 molecular weight cutoff ultrafiltration (panel B) of the same specimen. The general TIC patterns for both samples were somewhat similar, except for the broad peak at 38 min in the ACN-treated specimen. Panel C and D show the averaged mass spectra of species in 9 minutes of eluate (18-27 min) as indicated in panels A and B respectively. Comparison of panels C and D, shows that most of the peaks present in the ultrafiltrate were also present in the ACN-precipitated serum, but the ACN-treated specimen contained many additional molecular species, some quite abundant, that did not appear in the ultrafiltrate mass spectrum. This strongly supports the hypothesis that the ACN precipitation method liberated protein-bound species, making them available for cLC-MS analysis.

In order to confirm that ACN precipitation does indeed release and allow for analysis of more small proteins and peptides attached to serum albumin and other carrier proteins, additional experiments were carried out. The first of these involved the treatment of the large, retained proteins, present after ultrafiltration, with ACN to release potentially bound species. Figure 2.5 provides typical results from one set of experiments. In this approach the retentate obtained after a 30,000 molecular weight cutoff ultrafiltration of serum was explored by cLC-MS. The TIC of retentate prior to ACN precipitation (panel A) and after ACN precipitation (panel B) are shown. Panel C shows the averaged mass spectra overlay for MS data collected from 25 to 26 minute elution (as indicated in panels A and B). The retentate prior to ACN precipitation is shown in blue and the ACN-precipitated retentate is shown in red. As predicted, the retentate prior to ACN precipitation has a TIC spectrum very similar to that of unfiltered serum prior to ACN precipitation, and the mass spectrum is dominated by large molecular weight proteins. Prior ultrafiltration would have effectively removed unbound small proteins and peptides from the retentate. The appearance then of addition low molecular weight species after ACN precipitation then would evidence dissociation of bound small proteins and peptides from carrier proteins. This is strikingly evident in panel C which shows dramatically increased numbers of small molecular weight proteins and peptides after ACN treatment of the retentate and confirms that this approach makes

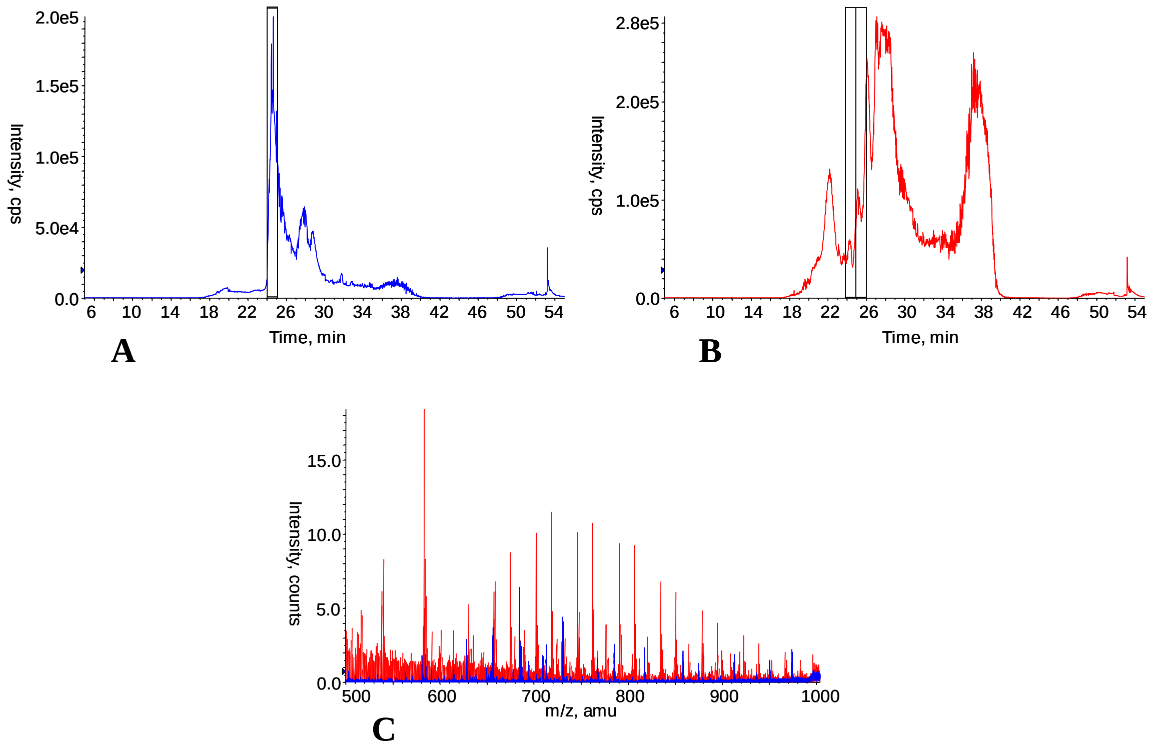


Figure 2.5. Examination of 30K NMW retentate before and after acetonitrile precipitation. Total ion chromatograms of 1 μg of ultrafiltration retentate prior to ACN precipitation (A) and retentate after ACN precipitation (B) are shown. C) The averaged mass spectra overlay for species present in eluate from 25 to 26 min of the cLC separation step are shown. Retentate prior to ACN precipitation is shown in blue, ACN-precipitated retentate is shown in red. The presence of the high abundance proteins masked the information rich presence of small peptides.

available many more small, low-abundance proteins and peptides for proteomic analysis. When these results were extended to a longer time region of the cLC chromatogram, similar results were obtained (data not shown).

An additional approach is summarized in Figure 2.6 which compares the mass spectra of ACN-precipitated retentate with the untreated filtrate obtained after ultrafiltration of a single serum. Panels A and B show the TIC of the ACN-precipitated retentate and the filtrate respectively. Panel C displays the averaged mass spectra overlay of 9 minutes of eluate (18-27 min) of the ACN-precipitated retentate (blue) and the filtrate (red). This overlay shows that ACN treatment produced many additional molecular species. Interestingly, while both procedures yield small proteins and peptides,

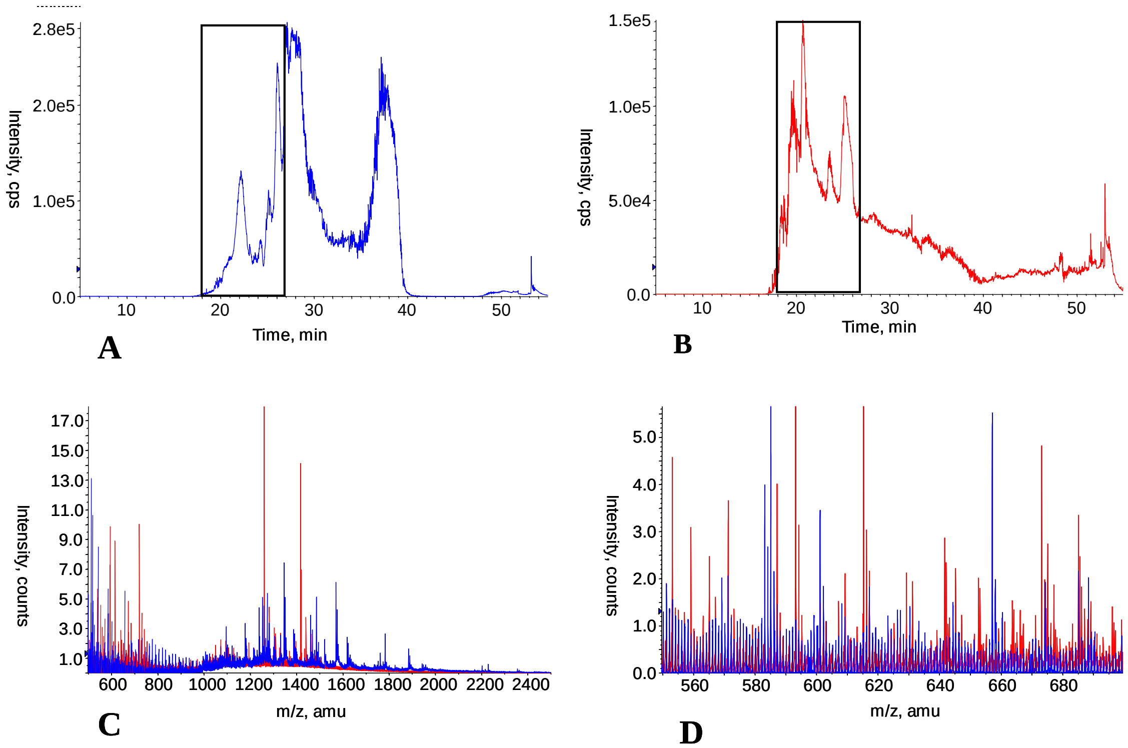


Figure 2.6. Comparison of ACN-precipitated retentate and filtrate. Total ion chromatograms of 1 µg of ACN-precipitated retentate (A) and 1 µg of filtrate (B) from 30 K NMW ultrafiltration are shown. C) The averaged mass spectral overlay for 18 to 27 min of the cLC chromatogram is presented. ACN-precipitated retentate is shown in blue; filtrate is shown in red. D) An expanded region (550 to 700 m/z) of Panel C is shown.

the MS spectrum from the ACN precipitate differed significantly from that of the filtrate, indicating that two subsets of small proteins and peptides exist and that those in the ACN-precipitated retentate were not representative of those that routinely passed through the filter.

Reproducibility

Not only is reproducibility in the precipitation step important, but reproducibility through the entire cLC-MS analysis is necessary for useful diagnostic applications of serum proteomics. The reproducibility of this method was tested. The comprehensive approach of ACN treatment with cLC-MS analysis was carried out on serum specimens

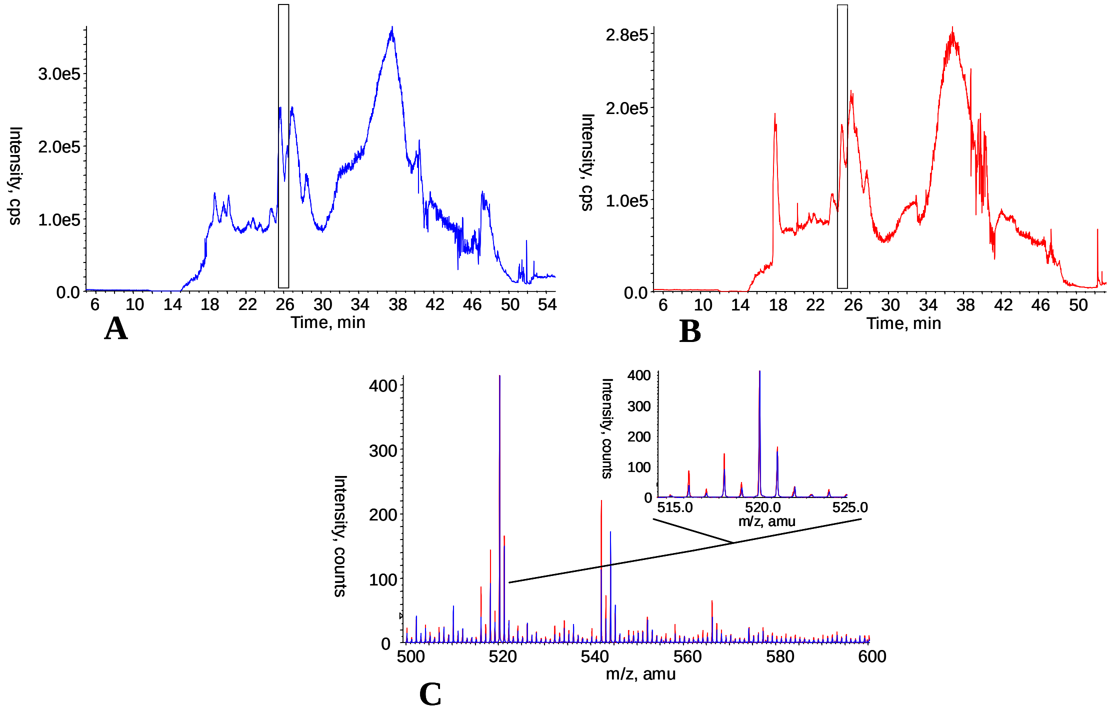


Figure 2.7. **Reproducibility of LCMS day to day.** Total ion chromatograms of ACN-precipitated serum of 2 different healthy subjects are shown. A) A TIC of subject 1 serum (blue), prepared and run on day 1 is shown. B) A TIC of subject 2 serum (red), prepared and run on day 2 is shown. C) An expanded mass spectral overlay (500-600 m/z) with inlay expanded to 550-560 m/z of both patients from a 1 min cLC window that had been calibrated for variations in elution time is shown (blue 25.4-26.4 min, red 24.8-25.8 min).

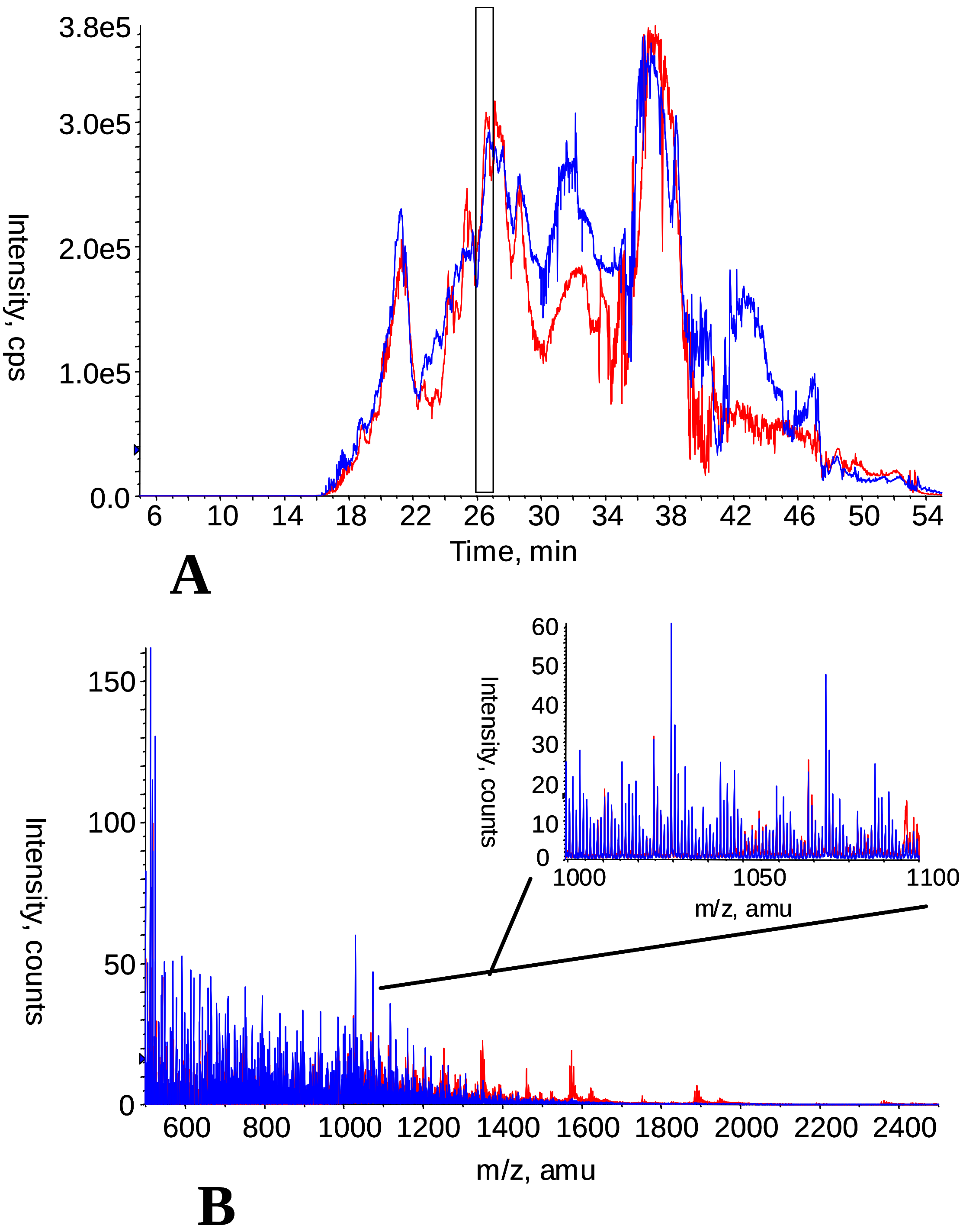


Figure 2.8. **Reproducibility of LCMS with variable sample handling.** A) An overlay of the TIC of ACN-precipitated serum thawed just prior to analysis (blue) with the TIC of serum left at room temperature overnight (red) is shown. B) An overlay of the averaged mass spectrum for cLC elution time of 26-27 min of the ACN-precipitated fresh serum (blue) with the averaged mass spectrum (same retention time) of the room temperature incubated serum (red) is provided. The inlay in panel B shows an expanded mass spectrum (1000 to 1100 m/z). Even with delayed handling of the serum, the mass spectra of the samples match closely

from a healthy patient, prepared and run on the same day. In addition comprehensive analysis of specimens from healthy patients prepared and run on different days was assessed.

Mass spectra obtained from different serum specimens from the same patient prepared and run on the same day were virtually indistinguishable (Data not shown).

Mass spectra obtained from different specimens from different healthy subjects prepared and run on different days were also remarkably similar after normalization of elution times (Figure 2.7).

The sensitivity of the results of our approach to variation in specimen handling was also explored (See Figure 2.8). The panel on the left (A) displays an overlay of the TIC of ACN-precipitated serum thawed just prior to use (blue) compared with the same specimen left at room temperature overnight (red). The panel on the right (B) shows the averaged mass spectra overlay of molecular species in one minute of cLC eluate (26-27 min) of the same freshly processed serum (blue) and serum processed after 24 hr at room

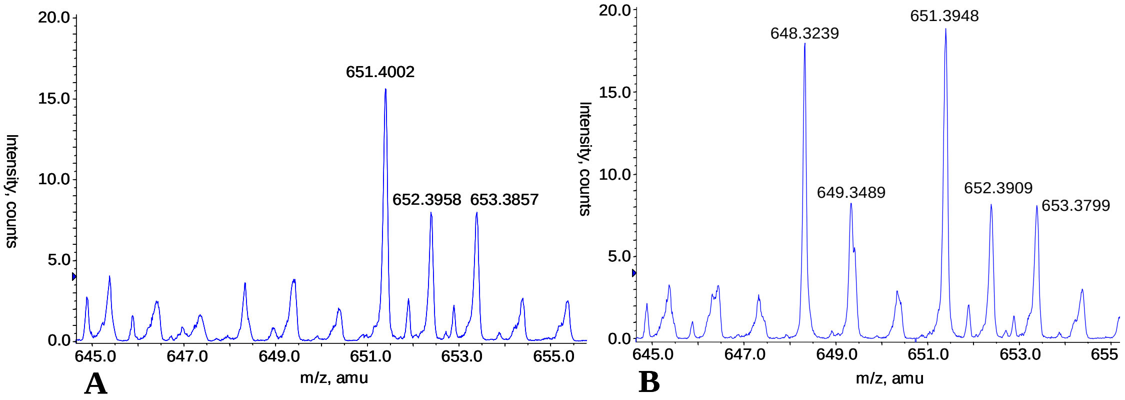


Figure 2.9: **Variability of LCMS from one individual to another.** The same retention time and m/z ranges were selected from cLC runs of a serum specimen of patient A (normal) and patient B (case). As shown, the variations in the make up of the serum from individual to individual can be detected. These differences represent a potential source of serum biomarkers.

temperature (red). Despite the delay, the mass spectra matched closely with modest changes.

Application of Method

As proof of principle of this approach, we applied this method to individuals representing two differing clinical states. Figure 2.9 shows a narrow region of the mass spectra (645-655 m/z) obtained after ACN treatment and cLC separation. A restricted region of the cLC chromatogram for each specimen was selected and an averaged mass spectrum developed for each. Panel A represents analysis of a specimen from a pregnant woman with an uncomplicated pregnancy. Panel B displays the results of a specimen from a pregnant woman of comparable gestational age who had preeclampsia, a severe complication of pregnancy. There was an additional molecular species at m/z = 648.3 in the preeclamptic specimen. This is one of several differences found. It is premature to know if the difference is consistently seen other preeclamptic patients. Nevertheless, it points out the very real potential of identifying differences between patient populations that might allow for the accurate diagnosis or prediction of disease, or even the eventual identification of mechanistic candidates.

Summary

In order to use the information-rich proteomic analysis of serum in a diagnostic manner, it is essential that the method used to prepare the sample provide reproducible results. The new method of acetonitrile precipitation described here adequately and reproducibly precipitates the non-informative, abundant, high molecular weight proteins. Moreover, this approach increases the number of low molecular weight species that can be analyzed by cLC-MS. We demonstrate importantly, that the new method of acetonitrile precipitation disrupts binding of small proteins and peptides from the large carrier proteins, providing a larger set of low molecular weight species for analysis. Most regulatory molecules have been shown to be small proteins and peptides and to be protein bound in serum; hence the possibility of developing a diagnostic pattern and identifying potential mediators of disease is increased using this approach. The combination of protein precipitation coupled with cLC-MS appears capable of meaningful evaluation of a substantial portion of the serum proteome.

Part II – Systematic Internal Standard Selection for Capillary Liquid Chromatography-Mass Spectrometry Time Normalization to Facilitate Serum Proteomics

Abstract

Because blood interacts with almost all tissues of the body, it is likely that changes in the overall health of an organism will be reflected in the quantities of specific serum peptides and proteins, making them biomarkers. Due to the complexity of serum, pre-analytical sample simplification and separation are needed prior to mass spectrometric analysis. Use of a reverse-phase capillary column coupled to a mass spectrometer allows for separation and analysis of serum as part of efforts to discover biomarkers. Even after sample simplification by organic solvent precipitation, data files for a single sample typically exceed 1 gigabyte, making it difficult to analyze serum MS profiles with currently available software. However, with adequate safeguards, it appears possible to find differences in peak intensities between clinical comparison groups visually. To facilitate this, an approach was developed where the elution profile was divided into 2 min intervals in which MS data were averaged. This required that molecular species had defined, reproducible elution times. Peaks generally had chromatographic elution peak widths of 1-1.5 minutes. Given cLCMS variation, misalignment of elution times of individual peaks occurred often. Hence, internal time controls were identified within each window and used for elution time normalization. The first has been identified as fibrinopeptide A. These species allowed for peak alignment across samples improving biomarker discovery.

Introduction

Medical diagnosis has benefited from the measurement of serum molecules that are present in lesser or greater abundance in individuals having a specific disease. It is anticipated that additional biomarkers could provide even more accurate diagnosis, including sub-classification of disease, staging of disease progression or even providing risk assessment as part of medical evaluation. If one could achieve a broad survey of biomolecules altered by a given disease, it may be possible to define molecules that actually participate in the disease mechanism and are consequently potential drug targets. Finding such markers then is of significant medical interest. Mass spectrometry (MS) has been gaining popularity as a means of searching for novel biomarkers in readily available bodily fluids such as serum, plasma (Hu 2006, Meng 2007, Drake 2007), urine (Mischak 2007), ocular fluids (Grus 2007, Wu 2007), and nipple aspirate fluid (Ruhlen 2007). Liquid chromatography coupled to mass spectrometry (cLC-MS) has been favored over other techniques in the search for new biomarkers in our lab due to its ability to provide a more comprehensive assessment of the peptides and small proteins present in complex samples than is possible with most other MS methods. This very ability to study hundreds to thousands of molecular ions in a single sample also generates large amounts of data, and so poses a major challenge to the use of such methods in finding new biomarkers due to increased complexity of data analysis. For example, a single cLC-MS run of a single serum sample typically produces an MS spectral data file that can reach one gigabyte or greater in size with 4000-5000 observable peaks. As one considers comparisons of clinical groups, each represented by dozens of specimens, the problem is compounded. One major challenge is having a straightforward way of addressing day-to-day or even run-to-run chromatographic variability so as to then allow direct comparisons of the spectra of samples run at different times.

Serum is an ideal medium in which to search for new biomarkers because it interfaces with the vast majority of cells in the body and so likely contains the imprint of diseased cells in the form of cell specific proteins and peptides, alterations in small signaling molecules, increased protein fragments, etc. (Zhang 2007) Also, because it is routinely collected at doctor visits, it is considered to be a minimally invasive specimen, and one that could be obtained repeatedly as part of prospective studies. Several groups have studied serum and plasma with the goal of proteomic biomarker discovery for the diagnosis of various cancers and other diseases. A recent review provides a good overview of putative biomarkers for human cancer discovered by serum proteome analysis. (Li 2002) Serum, though easily acquired and well suited for biomarker discovery, presents challenges. For example, serum contains tens of thousands of different polypeptides ranging in size from a few to several hundred amino acids and spanning nine orders of magnitude in concentration. (Adkins 2002, Anderson 2002) The 22 most abundant proteins in serum account for 99% of the total protein in serum. (Meng 2007)

Due to the extreme complexity of the protein complement found in serum, made even more complicated by variable post-translational modifications, including protein glycolsylations, a separation step is required. Capillary column chromatography allows for the direct interfacing of columns with the mass spectrometer while providing a robust fractionation step. However, whole serum can easily foul capillary columns. Hence, removal of abundant, but typically uninformative, proteins is desirable. Additionally, MS instrumentation typically has a detection dynamic range of only two orders of magnitude, so the observation of low abundance species is difficult without removal of the highly-abundant serum proteins. High-abundance species can cause signal suppression that blocks detection of lower abundance species. For all of these reasons, pre-analytical sample simplification of serum is needed prior to MS analysis.

A common method of simplifying a serum sample involves precipitating the large, high abundance proteins out of solution using an organic solvent. In addition to removing large proteins, this treatment has the added benefit of denaturing proteins acting as carriers, liberating smaller bound molecules. Because small proteins and peptides will be filtered from blood by the kidney (Adkins 2002, Liotta 2003, Petricoin 2002 Clinical), small molecules that the body deems important enough to keep around are bound to larger proteins to prevent their loss, giving them a longer half-life within the blood stream. Consequently, denaturing serum substantially increases the number of small proteins and peptides available for MS assay. However, many published methods of serum simplification remove these larger molecules without taking steps to ensure that any smaller proteins bound to them are released. It is in this low-molecular weight low-abundance fraction of the human serum proteome that many biomarkers are likely to be found (Meng 2007, Liotta 2003). Even with over 99% of total protein removed from the serum samples via organic solvent precipitation, there remain in solution thousands of molecular species. We have developed a cLC-MS proteomics approach to search for potential biomarkers in large and complex data sets produced from analysis of human serum samples. Protein removal is followed by a reverse-phase capillary column step to help to separate the sample over time, reduce signal-suppression and allow us to get a better overall picture of the state of the fraction of the low-molecular weight proteome of the patient that we are focusing on.

When such a sample is analyzed using cLC-MS, a large and very complex three dimensional data set is produced. Currently available software has been very good at three dimensional comparisons of simple data sets, but has proven inadequate when working with the large, complex data sets produced using our methods. Lack of current software capable of analyzing these data sets has led to searching for peptides of interest within our data manually.

The complex nature and sensitivity of the LC instrumentation employed, as well as slight variations in the hand-packed columns used caused the elution times of peptides and proteins to vary from day-to-day and run-to-run. Although the start time of the overall elution profile may vary from sample to sample, the elution order and the relative elution times of molecular species within a single sample run stay generally constant. Because most chromatographic peaks require 1-2 min to elute completely and in an effort to reduce the size of data sets and keep peak number manageable, we defined 2 minute elution windows along the chromatogram to allow comparison across several specimens in a sample set or between sample sets. Within each window we located a central peak, typically a peptide, found in all specimens that could be used as a reference allowing for time alignment. Once aligned, the spectra can be overlaid and searched for quantitative differences between comparison groups. Any differences discovered can then be further studied as possible biomarkers.

Materials and Methods

All reagents were purchased from Sigma (St. Louis, MO) and used without further purification.

Sample Collection

Blood was obtained by antecubital venipuncture at the BYU Student Health Center from a healthy female volunteer. Blood was allowed to clot for 30 minutes at room temperature, and after centrifugation at 3,500 rpm for 15 min, serum was collected, aliquoted and frozen and maintained at -80°C until further processing.

Acetonitrile Precipitation

This method has been described in detail in a previous paper. (Merrell 2004) Briefly, two volumes of HPLC grade acetonitrile (400 µl) were added to 200 µl serum, vortexed vigorously for 5 sec and allowed to stand at room temperature for 30 min. Samples were then spun for 10 min at 12,000 rpm in a centrifuge at room temperature. An aliquot of supernatant (550 µl) was then transferred to a new microcentrifuge tube to which was added 300 µl HPLC grade water, the two mixed, and the mixture then lyophilized to ~200 µl in a vacuum centrifuge (LABCONCO CentriVap Concentrator) to remove any acetonitrile. Supernatant protein concentration was determined using a Bio-Rad microtiter plate protein assay (BioRad, Hercules, CA) according to manufacturer instructions. An aliquot containing 4 µg protein was transferred to a new microcentrifuge tube and lyophilized to a volume < 20 µl. The sample was brought to 20 µl with HPLC water to which was added 20 µl 88% formic acid and the sample was mixed vigorously.

cLC-MS Analysis

Capillary liquid chromatography (cLC) to fractionate or separate peptides and proteins was performed using a 15 cm x 250 um i.d. capillary column, packed in-house using POROS R1 reversed-phase media (Applied Biosystems, Framingham MA).

Elution was accomplished employing a 2.2%/min gradient from 0% organic to an organic concentration of 60% acetonitrile in 0.1% formic acid, followed by a 3.5%/min gradient up to a concentration of 95% organic phase. In-line chromatographic separation used an LC Packings Ultimate Capillary HPLC pump system with a FamOS autosampler (Dionex Corp) controlled by the mass spectrometer software (Analyst, Applied Biosystems). The cLC was coupled directly to the MS. Effluent from the capillary column was directed into a QSTAR Pulsar i quadrupole orthogonal time-of-flight mass spectrometer through an IonSpray source (Applied Biosystems). Data was collected for m/z 500 to 2500 over the entire chromatogram (55 min total including void volume, elution and reequilibration). Data collection, processing and preliminary formatting were accomplished using the Analyst QS software package with BioAnalyst add-ons (Applied Biosystems).

cLC-MS/MS Analysis

A specimen containing 0.5 μg of total protein was loaded onto the column and a MS run was performed to determine the exact elution time of the peptide of interest. A MS/MS run was performed on the same amount of protein with collection of Multi-Channel Acquisition (MCA) fragmentation data at a set collision energy of 30 for the two min time span in which the peptide of interest eluted. The fragmentation spectra were sent to MASCOT for identification. The first peptide used for elution time normalization had a mass of 1464.6 and was successfully identified as fibrinopeptide A.

Selection of Reference Peaks for Chromatographic Normalization

Using 2-D and 3-D visualizations of complete serum runs, multiple peaks that eluted at ~2 min intervals were selected as possible reference peaks. The Extract Ion

Chromatogram (XIC) function was used to check the chromatographic elution profile of a small m/z range that included the m/z of the possible reference peaks. To be further considered as possible reference peaks, chromatographic elution profiles of the selected peaks had to be relatively narrow (< 2 min), well shaped (close to Gaussian in shape), and having elution profiles that were well resolved and distinguishable from those of other molecular species of similar m/z. Those peaks that best met these criteria were further investigated in serum runs from various individuals prepared and run on different days. Ten peaks that eluted at ~2 min intervals and were ubiquitously present in all serum samples were selected as markers.

Normalization of Elution Times

Differences in elution time were corrected by aligning the central reference peaks. In practice this was accomplished by generating an averaged mass spectrum of the interval represented by exactly1.00 min before and after a given reference peak, irrespective of its actual chromatographic elution time. Comparable mass spectra were generated for each specimen in the comparison groups for each of the 10 internal time controls. To evaluate the utility of the approach 2 peaks (one running before and running after the reference peak for first window) were monitored in several samples of a single pool of normal serum. There were several aliquots of this pool assayed on the same day (n=10 on two separate occasions) and several aliquots (n=17) of the pool assayed over several days. Initially, the actual elution time of each test peak was obtained from the XIC plot of each of 27 runs, 10 run on the same day and this done twice and 17 on other days. As a quantitative indication of the effectiveness of time normalization we looked at variability around two means. The first mean was the mean elution times of these two peaks as established for the whole set, as well as for same day and day-to-day runs. Then the mean elution time variance was calculated as was the absolute difference between that mean and an individual elution time for each analysis. These differences between actual value and mean value were considered the measure of variability and were compared with normalized data. To determine variability of the time normalized data first the time difference between the internal control and each of the two test peaks for each sample was calculated. This is termed the offset time. These offset times in turn were averaged and the absolute difference between a given offset time and the mean was calculated. The variability in this difference was then considered to be the variability after normalization and was compared to the un-normalized data statistically.

Statistical Analysis

Numerical data are presented as the mean ± SE of the mean. Comparisons of variability were made by compiling the individual elution times for test peaks, determining the mean and finding the absolute difference between an individual elution time and the mean. As described about the variability in the difference between the mean elution time and the actual individual elution times was then compared to normalized data. For the normalized data the same central internal control was used, the same test peaks were used and their time difference (offset time) to the internal control was determined. These values were averaged and the variability in offset times was estimated by calculating the difference between the mean offset time and each individual offset time. The two sets of time variability were compared by Student's t-test. A p-value <0.05 was considered significant.

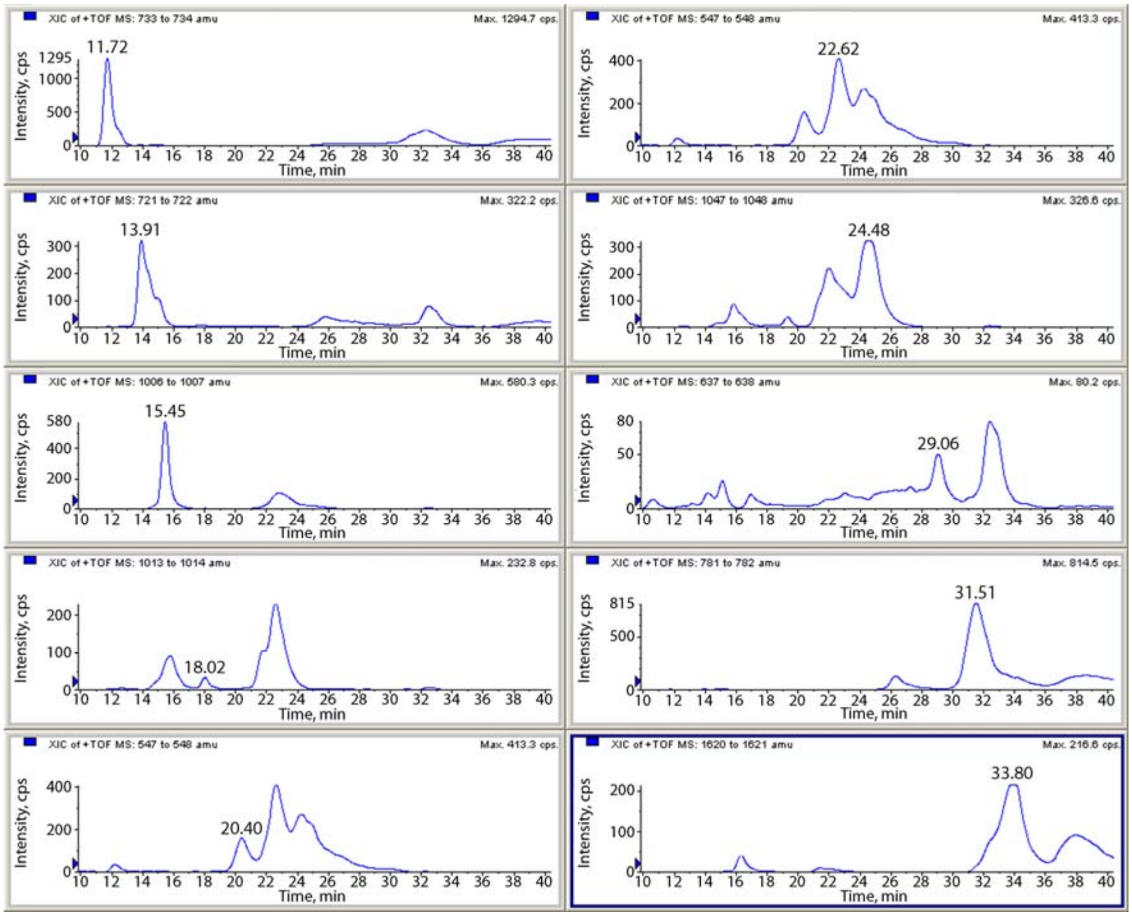


Figure 2.10. Selection of peptide elution standards. Labeled peaks within the chromatograms shown represent the XIC elution profiles of the 10 molecular species chosen as standards.

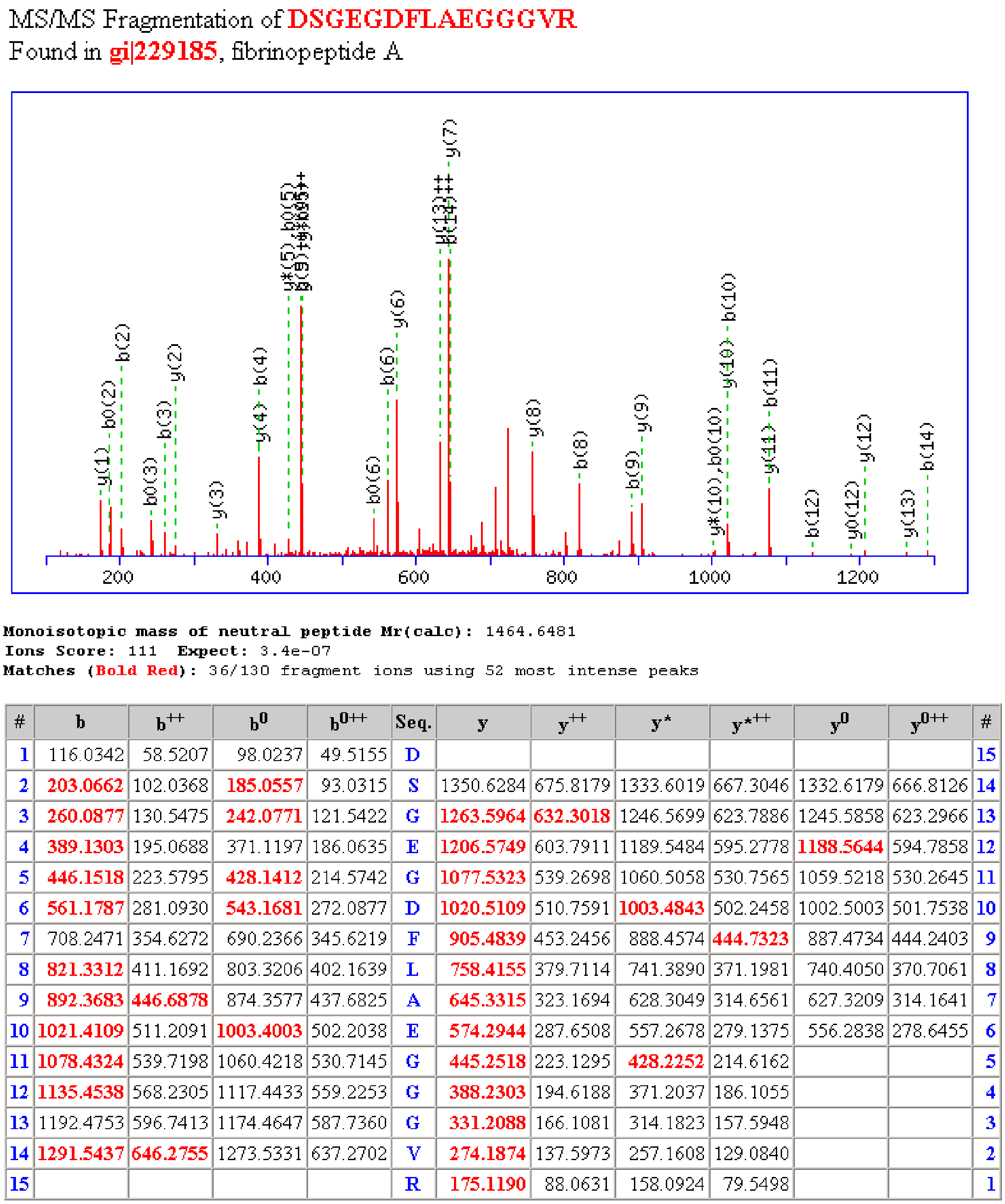
Results

In an effort to correct for differences in day-to-day and even run-to-run chromatographic variation between samples an approach was developed using internal controls. To do this 10 molecular ion peaks were selected that were present in all human serum samples and that eluted at approximate 2 minute intervals throughout the elution period. These were used for time normalization of serum samples run by cLC-MS (Figure 2.10) It was unimportant as to whether the peaks selected were in any particular charge state, whether they represented the monoisotopic peak within a charge envelope, or were even peptidic in nature. (Table 2.2) Provided that these internal control species had a recognizable m/z value, eluted at a point of interest, were consistently found in all specimens and were chromatographically well behaved, their structure or chemical identities were not necessary for them to be useful as a reference. In reality, the identities of these control species are unlikely to be interesting in the context of biomarker discovery, given their abundant and uniform nature.

Marker #	**m/z of reference peak**	**Charge State**	**Is monoisotopic**	**Relative Elution Time (min)**	**XIC Window**
1	**733.3**	**+2**	**Yes**	**0**	**733-734**
2	**721.3**	**+2**	**No**	**2.2**	**721-722**
3	**1006.0**	**+2**	**Yes**	**3.7**	**1006-1007**
4	**1013.5**	**+5**	**Yes**	**6.3**	**1013-1014**
5	**547.3**	**+1**	**Yes**	**8.7**	**547-548**
6	**547.3**	**+1**	**No**	**10.9**	**547-548**
7	**1047.7**	**+1**	**Yes**	**12.8**	**1047-1048**
8	**637.3**	**+1**	**Yes**	**17.3**	**637-638**
9	**781.5**	**+1**	**No**	**19.8**	**781-782**
10	**1620.2**	**+1**	**Yes**	**22.1**	**1620-1621**

Table 2.2. The m/z value of the selected ten reference peaks is shown along with their charge state, position in the isotope envelope, average elution time (n=160) relative to that of the first standard, and the values used for each XIC window.

Nevertheless, once the 10 alignment controls were chosen as markers, attempts were made to sequence those that were peptides using MS/MS. These efforts provided partial but insufficient amino acid sequence data to allow identification of all but one of the standards due to their large size (m/z 1013.5), the species being non-peptidic in nature (m/z 547.3), fragmentation being poor (m/z 1006.0), fragmentation producing only very small fragments (m/z 721.3), or for other reasons. To date only the first standard (m/z 733.3 where z = 2) has been successfully identified. The peptide was Fibrinopeptide A (Figure 2.11).



MS/MS Fragmentation of DSGEGDFLAEGGGVR
Found in gi|229185, fibrinopeptide A

```
Monoisotopic mass of neutral peptide Mr(calc): 1464.6481
Ions Score: 111  Expect: 3.4e-07
Matches (Bold Red): 36/130 fragment ions using 52 most intense peaks
```

#	b	b++	b0	b0++	Seq.	y	y++	y*	y*++	y0	y0++	#
1	116.0342	58.5207	98.0237	49.5155	D							15
2	**203.0662**	102.0368	**185.0557**	93.0315	S	1350.6284	675.8179	1333.6019	667.3046	1332.6179	666.8126	14
3	**260.0877**	130.5475	**242.0771**	121.5422	G	**1263.5964**	**632.3018**	1246.5699	623.7886	1245.5858	623.2966	13
4	**389.1303**	195.0688	371.1197	186.0635	E	**1206.5749**	603.7911	1189.5484	595.2778	**1188.5644**	594.7858	12
5	**446.1518**	223.5795	**428.1412**	214.5742	G	**1077.5323**	539.2698	1060.5058	530.7565	1059.5218	530.2645	11
6	**561.1787**	281.0930	**543.1681**	272.0877	D	**1020.5109**	510.7591	**1003.4843**	502.2458	1002.5003	501.7538	10
7	708.2471	354.6272	690.2366	345.6219	F	**905.4839**	453.2456	888.4574	**444.7323**	887.4734	444.2403	9
8	**821.3312**	411.1692	803.3206	402.1639	L	**758.4155**	379.7114	741.3890	371.1981	740.4050	370.7061	8
9	**892.3683**	**446.6878**	874.3577	437.6825	A	**645.3315**	323.1694	628.3049	314.6561	627.3209	314.1641	7
10	**1021.4109**	511.2091	**1003.4003**	502.2038	E	**574.2944**	287.6508	557.2678	279.1375	556.2838	278.6455	6
11	**1078.4324**	539.7198	1060.4218	530.7145	G	**445.2518**	223.1295	**428.2252**	214.6162			5
12	**1135.4538**	568.2305	1117.4433	559.2253	G	**388.2303**	194.6188	371.2037	186.1055			4
13	1192.4753	596.7413	1174.4647	587.7360	G	**331.2088**	166.1081	314.1823	157.5948			3
14	**1291.5437**	**646.2755**	1273.5331	637.2702	V	**274.1874**	137.5973	257.1608	129.0840			2
15					R	**175.1190**	88.0631	158.0924	79.5498			1

Figure 2.11. Identification of the first peptide standard. The first serum standard has been identified as Fibrinopeptide A using cLC-MS/MS. The fragmentation spectra and resultant identification made by submitting the data to the MASCOT search engine are shown above.

Time windows (2.0 min) centered on each of the XIC elution times of the ten standards allowed for time normalization. This in turn allowed for the overlay of up to 16 different spectra averaged over that 2 min window. As a demonstration of the utility of this approach several separated runs of a single human serum (limiting biologic variability) were carried out. Two peaks in the first time window were selected having m/z ratios of 675.4 and 1050.5 respectively. The actual elution time for each of these two species was determined for 37 separate samplings. Then the relevant windows were aligned and the time differences between runs were calculated from the reference peak (m/z 733.4). This brought about a marked and significant reduction in the variability of the time measurement (Table 2.3). Specimens run on the same column, same day were more consistent in terms of the actual elution times than those run on separate days. Nevertheless, the internal time controls allowed for a significant reduction in variation in either case (Table 2.3).

Discussion

Ten internal serum peptides were selected as standards for cLC-MS time normalization. Many more peaks met the criteria needed to be considered as internal standards. Those chosen were selected in part for their location within the cLC chromatogram and their convenience in use. It was unimportant as to whether the peaks selected were in any particular charge state, whether they represented the monoisotopic peak within a charge envelope, or were even peptidic in nature.

Of those finally selected most were monoisotopic, but a few were not. The MS instrumentation is sensitive enough to detect sub-dalton differences in mass. Due to the natural abundance of ^{13}C, any molecule containing carbon atoms has a chance of one or more of those carbons being ^{13}C instead of the more common ^{12}C. Molecules detected using MS typically show up as a series of peaks, representing the mass change that results from addition of one, two, or more neutrons as a result of incorporation of ^{13}C molecules.

Data set	**No normalization**	**Normalized**	
Same day set 1 (n=10)	From mean ± SD	From mean ± SD	P
Peak 1	0.090 ± 0.048 min	0.010 ± 0.014 min	0.00029
Peak 2	0.060 ± 0.040 min	0.010 ± 0.012 min	0.0016
Same day set 2 (n=10)	From mean ± SD	From mean ± SD	P
Peak 1	0.110 ± 0.092 min	0.020 ± 0.017 min	0.0070
Peak 2	0.080 ± 0.065 min	0.010 ± 0.013 min	0.0044
Day-to-day set 1 (n=17)	From mean ± SD	From mean ± SD	P
Peak 1	1.06 ± 0.88 min	0.15 ± 0.075 min	0.000175
Peak 2	0.90 ± 0.84 min	0.23 ± 0.152 min	0.003
All determinations (n=37)	From mean ± SD	From mean ± SD	P
Peak 1	0.78 ± 0.65 min	0.09 ± 0.07 min	0.002
Peak 2	0.80 ± 0.57 min	0.18 ± 0.11 min	0.003

Table 2.3. Comparison of variability in raw elution times without time alignment with the variability in offset times after time alignment for two peaks. Peak 1 = m/z 675.4, peak 2 = m/z 1050.5. Both peaks were found in the first elution window which also contained the central time alignment peak (m/z 733.4).

The peak that represents the molecule with all ^{12}C and no ^{13}C is called the monoisotopic peak. If a monoisotopic molecular species did not conform to the selection criteria, a species within the ^{13}C isotope envelope that did conform to the selection criteria was chosen instead. Use of these internal peaks as controls produced consistent elution time alignment across all sample runs and this in turn allowed for visual inspection of spectral overlays in the search for potential biomarkers. It is worth noting that a highly specific and useful proteomic control does not require knowledge of its actual composition provided it has a consistent and identifiable mass, is sufficiently abundant and resolved from other species. Attempts at identification of the ten time standards have found that some of the current markers are not peptides, but they have still proven useful in this approach. Likewise, a highly specific proteomics pattern can be useful in the prediction or diagnosis of a disease in the absence of the complete knowledge of the chemical composition of the individual constituents in that pattern, although such information may prove useful in the understanding of the disease process.

Using this approach, we demonstrated that normalizing elution times in the straight forward manner described here markedly decreased timing misalignment for any specific peak across multiple analyses of the same specimen. This was evident as a marked decrease in variability of elution time relative to the standard versus the actual cLC elution times. The impact of this is potentially large. For intraassay variability the average deviation from the mean was ~0.1 min without normalization but some specimens differed by as much as 0.3 min on either side. Considering the 2 standard deviations, this creates a range of values that varies somewhat over 0.3 min. Even so, this is reduced to ~0.01-0.02 min with a range of values that spans 0.04 min after normalization. The intraassay variability is even more impressive. There the mean elution time variability was ~ 1 min with a range (± 2 SD) of ~ 3.5 min. After time alignment the mean time variability was ~ 0.2 min with a range of 0.4 min. Given that our elution windows are 2.0 min, it is entirely possible that a given peak might be inadvertently omitted if time normalization were not employed.

The significance of these results is primarily in the context of biomarker discovery. With advances in mass spectrometric technology has come the possibility of assessment of biological molecules, even in the serum. Likewise, with more innovative proteomic approaches there have been and will continue to be dramatic increases in both the size and complexity of the data sets accessed by the instrumentation. The need for additional biomarkers for medical diagnosis and clinical risk assessment of many diseases is widely acknowledged. Analysis of these data is becoming increasingly difficult and advances in software capable of analyzing these large and complex data sets has lagged behind advances in instrumentation. cLC-MS produces such massive data sets that current software cannot manage such file size and it is necessary to parse the elution interval into discrete reproducible windows. While this approach to data analysis reduces the file size and makes more manageable the absolute number of peaks to be

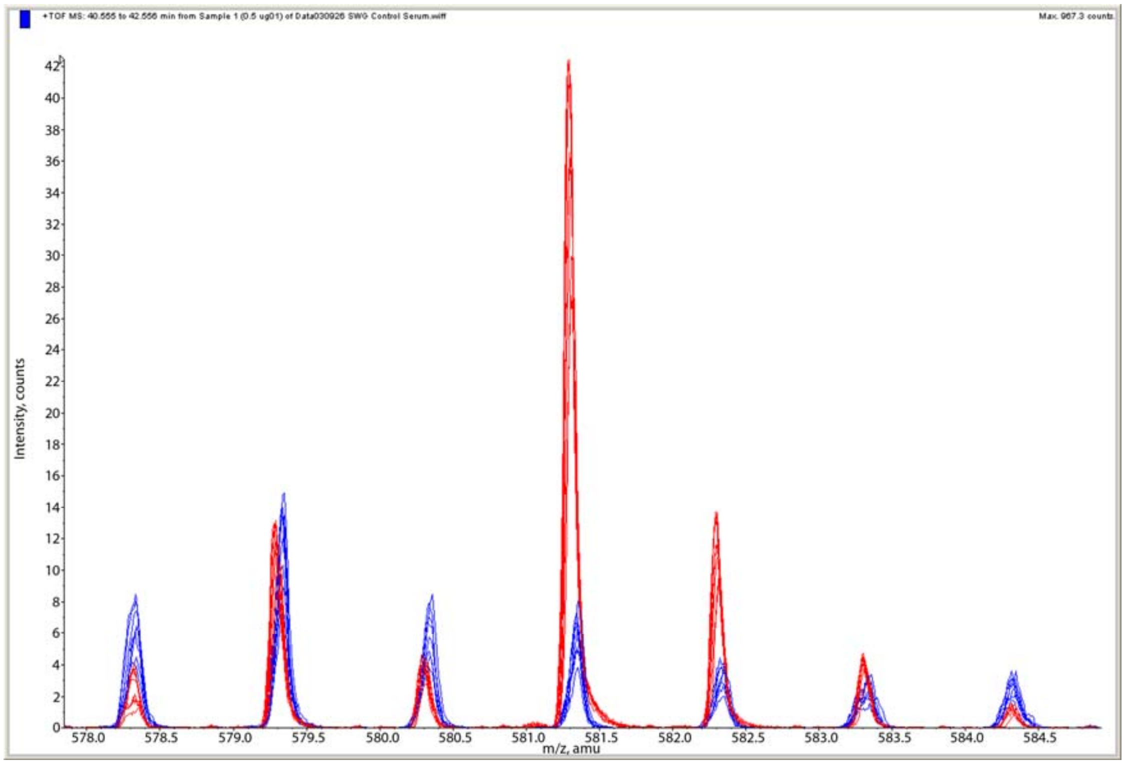


Figure 2.12. Spectra overlay for biomarker discovery. The XIC elution time for one of the standards was recorded for each of 8 male (blue) and 8 female (red) sample runs. A two-minute window was centered around this time and the resultant spectra were overlaid. The resultant spectral overlay can then be searched for biomarker candidates. Notice that in this case the intensity of the blue peaks was greater than the red for most of the spectral overlay, but the pattern suddenly changed for the peaks in the middle. Such differences could be recorded and the peaks would be further investigated as possible biomarkers.

considered, it creates the possibility that peaks of interest actually overlap two windows or to a substantial degree fall outside the relevant time window. If time is misaligned, then more or less of a given molecular species will be included in the analyzed window potentially giving rise to differences in peak heights and areas due to location within the window, which is related to the uncertainties in elution time variability rather than clinically related differences. When spectra from two clinical groups are overlaid, our experience has clearly shown that the eye is very good at locating peaks that are quantitatively different. (Figure 2.12) However, to insure that the quantitative differences are due to biological change and not to the partial inclusion or exclusion of a peak due to chromatographic time differences, the use of this time alignment approach involving endogenous internal controls dramatically reduces this problem and allows for successful identification of peaks that differ between clinical groups. While this approach is somewhat laborious and time intensive, use of this method has led to the discovery of several statistically significant novel clinical biomarkers that are under active investigation.

Part III – A comprehensive serum proteomic approach capable of monitoring the low molecular weight proteome with sequencing of intermediate to large peptides.

Summary

The low-abundance, low molecular weight serum proteome has high potential for new biomarker discovery using mass spectrometry (MS). Because the serum proteome is large and complex, defining quantitative differences for a molecular species between comparison groups requires an approach with robust separation capability, high sensitivity, as well as high mass resolution. Capillary liquid chromatography (cLC)-MS provides both the necessary separation and the sensitivity to observe thousands of low-abundance peptides. Subsequent identification of potential serum peptide biomarkers observed using cLC-MS can in principle be accomplished by cLC-MS/MS without further sample preparation or additional instrumentation. In this report a novel cLC-MS/MS method for peptide sequencing is validated that surpasses previously reported size limits for amino acid sequencing accomplished by fragmentation studies.

As a demonstration of the approach, two low-abundance peptides with a mass of ~ 4000–5000 Da were selected for MS/MS identification. The multi-channel analyzer (MCA) was used in a novel way that allowed for summation of 120 fragmentation spectra per each of multiple customized collision energies, providing more thorough fragmentation coverage of the peptide with improved signal to noise. The peak list from this composite analysis was submitted to Mascot for identification. The index 4279 Da and 5061 Da peptides were successfully sequenced and identified from a single sample using cLC-MS followed by cLC-MS/MS on a single platform. The peptides were a 39 amino acid immunoglobulin G heavy chain variable region fragment and a 47 amino acid

fibrin alpha isoform C-terminal fragment. The methods described here provide the ability on a single platform to both survey thousands of serum molecules and couple that with a dramatic enhancement in the peptide sequencing capabilities of cLC-MS/MS , providing an effective technique for serum biomarker discovery.

Introduction

Serum is a promising source from which to identify novel biomarkers. Blood is intimate with most regions of the body, and it is likely that changes accompanying abnormal or disease states could be reflected in the low molecular weight complement of the blood. In addition, because serum can be collected in a minimally invasive way, it is more easily and acceptably available for analysis than other tissues. However, serum is complex and several challenges are involved in using it in the search for biomarkers. These challenges fall generally into three categories: the complexity of the specimens themselves, the limits of instrumentation, and the analysis of the data.

Because no serum proteomic method is able to measure all peptides and proteins, any serum proteomic approach requires the rational selection of which proteome cross-section will be assessed. When working with serum, the removal of bulk proteins from serum can be important for several reasons: Ninety-nine percent of the protein in serum is comprised of only 22 abundant proteins. (Anderson 2002) The concentration of the least abundant protein in serum differs from that of the most abundant by over 10 orders of magnitude. (Anderson 2002, Jacobs 2005) Additionally, capillary liquid chromatography separations of whole serum are problematic. Consequently, depleting the sample of bulk proteins is vital to observing the thousands of less abundant proteins and peptides.

Furthermore, the fraction of the serum proteome that has the highest potential for new biomarker discovery has been suggested to be the low-abundance, often lower molecular weight (LMW) proteins and peptides. (Liotta 2003) Because many regulatory peptides and proteins are bound to large carrier proteins in the blood, to study these compounds' role in health and disease, they need to be displaced from their larger carrier proteins.

All of these considerations make organic solvent precipitation of larger, serum proteins—including the abundant and less informative proteins—with its added ability to liberate bound species an attractive first step in accessing this important serum proteome subset. (Merrell 2004)

The complexity of serum (even of protein-depleted samples) places high demands on the analytical approach, especially the instrumentation. The four most common MS platforms used for proteomic analysis are matrix-assisted laser desorption/ionization mass spectrometry (MALDI) (Hortin 2006), Fourier-transform ion cyclotron resonance (FTICR) (Bogdanov 2005), surface-enhanced laser desorption/ionization (SELDI) (Li 2002, Diamandis 2004), and capillary liquid chromatography coupled to online electospray-ionization mass spectrometry (cLCMS) (Merrell 2004, Koomen 2005). Of these, only cLCMS provides a sufficiently robust separation capability to survey the thousands of peptides and LMW proteins present.

Once quantitative differences can be documented using serum proteomics, a protein or peptide is ideally sequenced prior to use as a potential biomarker. cLCMS facilitates this identification because of the seamless shift from cLCMS in the discovery stage to cLCMS/MS in the identification stage. Identification by cLCMS/MS has been achieved by either collisionally-induced dissociation (CID) or electron capture dissociation (ECD). In CID, repeated collisions of the peptide of interest with inert gas confer energy to the molecule, causing fragmentation at the weakest peptide bond. ECD does not involve a transfer of energy and so the resulting fragmentation does not depend on bond strength. (McLafferty 2001) Because of the differing mechanisms of fragmentation, these techniques are thought of as complementary, but a recent comparison of the two methods showed only a 12% overlap in identification between CID and ECD. Molecules with charge states above +2 were more effectively fragmented with ECD, while those with charge states of +1 and +2 were more effectively fragmented with CID. (Good 2007) An ideal method would utilize both of these identification strategies. Unfortunately, instruments capable of ECD are costly and relatively uncommon compared with instruments that use CID as their sole method of fragmentation. As proteins get larger, the efficiency of the CID process decreases with a smaller proportion of the backbone bonds cleaved. (Loo 1988, Horn 2000) Some peptide bonds are more labile than others, and thus more likely to be fragmented. (Zubarev 2008) As a result, not all fragments will be present at equal yield. Despite these complications, we show here that it is still possible to get useful fragmentation data from large, multiply-charged peptides using CID.

The approach described can overcome the significant challenges of CID-induced cLCMS/MS of peptides and LMW proteins including: 1) low signal-to-noise of daughter ions produced by fragmentation of peptides larger than 20 amino acids in length, 2) low dwell time at specific collision energies (CE) because of the instrument typically imposing rolling collision energies, and 3) the high complexity of the fragmentation data produced and submitted to database searches. As proof of principle, the comprehensive proteomics approach described here has been used to sequence two low-abundance index peptides, both having charge states of +6, and 39 and 47 amino acids in length respectively. These were not sequenced using the standard automated approach. The customized approach explained here then offers a way to extend the capability of cLCMS coupled with cLCMS/MS as a tool for peptide biomarker discovery.

Experimental Procedures

Sample Collection

Blood samples were obtained at the Brigham Young University Student Health Center from a healthy female volunteer, collected into serum separator tubes and processed after clotting by centrifugation to produce serum. The serum sample was aliquotted, frozen, and stored at -80° C until use.

Acetonitrile Precipitation

High molecular weight, high-abundance proteins were precipitated out of solution using a previously published method (Merrell 2004), modified as noted below. Briefly, two volumes of HPLC grade acetonitrile (400 µl) were added to 200 µl serum, the sample was vortexed vigorously for 5 sec and allowed to stand at room temperature for 30 min. Samples were then centrifuged for 10 min at 13,400 x g at room temperature. An aliquot of the supernatant (~550 µl) was then transferred to a microcentrifuge tube containing 300 µl HPLC grade water. The sample was vortexed briefly to mix and then lyophilized to ~200 µl in a vacuum centrifuge (Labconco CentriVap Concentrator, Labconco Corporation, Kansas City, MO). Water added prior to lyophilization aided in the complete removal of acetonitrile from the solution and is the only change from the previously published method. This step was necessary because acetonitrile is incompatible with the assay used to determine protein concentration and complicates sample loading onto the reversed-phase cLC column. Supernatant protein concentration was determined using a Bio-Rad microtiter plate protein assay performed according to manufacturer instructions (Bio-Rad, Hercules, CA). An aliquot of the same supernatant containing 4 μg of total protein was transferred to a new microcentrifuge tube and lyophilized to a volume less than 20 μl. Lyophilized samples were brought to 20 μl with HPLC water and acidified by addition of 20 μl 88% formic acid. The excess, unacidified supernatant was stored at –20° C for subsequent LC/MS analysis.

LC Separation

Protein-depleted, acidified serum samples (40 μl) were placed into a FAMOS autosampler 48 well plate (Dionex Corporation, Sunnyvale, CA) kept at 4° C. The autosampler injected 0.5 μg (5 μl) of each serum sample onto the guard column using HPLC water acidified with 0.1% formic acid at a flow rate of 40 μl/minute.

Capillary liquid chromatography used an LC Packings UltiMate Capillary HPLC pump system, with a SWITCHOS switcher and flow splitter, and a FamOS autosampler (Dionex), controlled by Analyst QS® software supplied with the QSTAR® mass spectrometer (Applied Biosystems, Foster City, CA).

The cLC columns included a 1 mm (16.2 μl) microbore guard column (Upchurch Scientific, Oak Harbor, WA) and a 15 cm x 250 um i.d. capillary separation column assembled in-house. The guard column was dry-packed and the capillary column was slurry packed using POROS R1 reversed-phase media (Applied Biosystems, Framingham, MA). Column equilibration and chromatographic separation were performed using gradient elution employing an aqueous phase (98% HPLC H_2O, 2% ACN, 0.1% Formic Acid) and an organic phase (2% HPLC H_2O, 98% ACN, 0.1% Formic Acid). Separation was performed beginning with a 3 minute column equilibration at 95% aqueous solution, followed by a 2.75%/min gradient to 60% organic phase, which was then increased at a rate of 7% increase/min to a concentration of 95% organic phase. The gradient was held at 95% organic phase for 4 minutes to elute the more hydrophobic portion of the sample, then dropped to 5% organic phase to re-equilibrate the column. All separations were performed at a flow rate of 5 μl/min.

MS Analysis

Effluent from the capillary column was directed into a QSTAR Pulsar i quadrupole orthogonal time-of-flight mass spectrometer through an IonSpray source (Applied Biosystems). Data was collected for m/z 500 to 2500 beginning at 5 and ending at 40 min elution time. Data collection, processing and preliminary formatting were accomplished using the Analyst QS software package with BioAnalyst add-ons (Applied Biosystems).

Peptide Selection

To choose a peptide for cLC-MS/MS identification, the cLC-MS spectrum was visually inspected for multiply charged peptides with a mass of 4000-5000 Da that were present in relatively low amounts compared with other molecular species eluting near the same time. The selected peptides needed to have eluted completely within a two-minute window and needed to have good chromatographic separation from other molecular species of similar m/z. The peptides that were chosen for this proof-of-principle study had masses of 4279.2 Da (m/z 714.2) and 5061.1 Da (m/z 844.6), with relative average

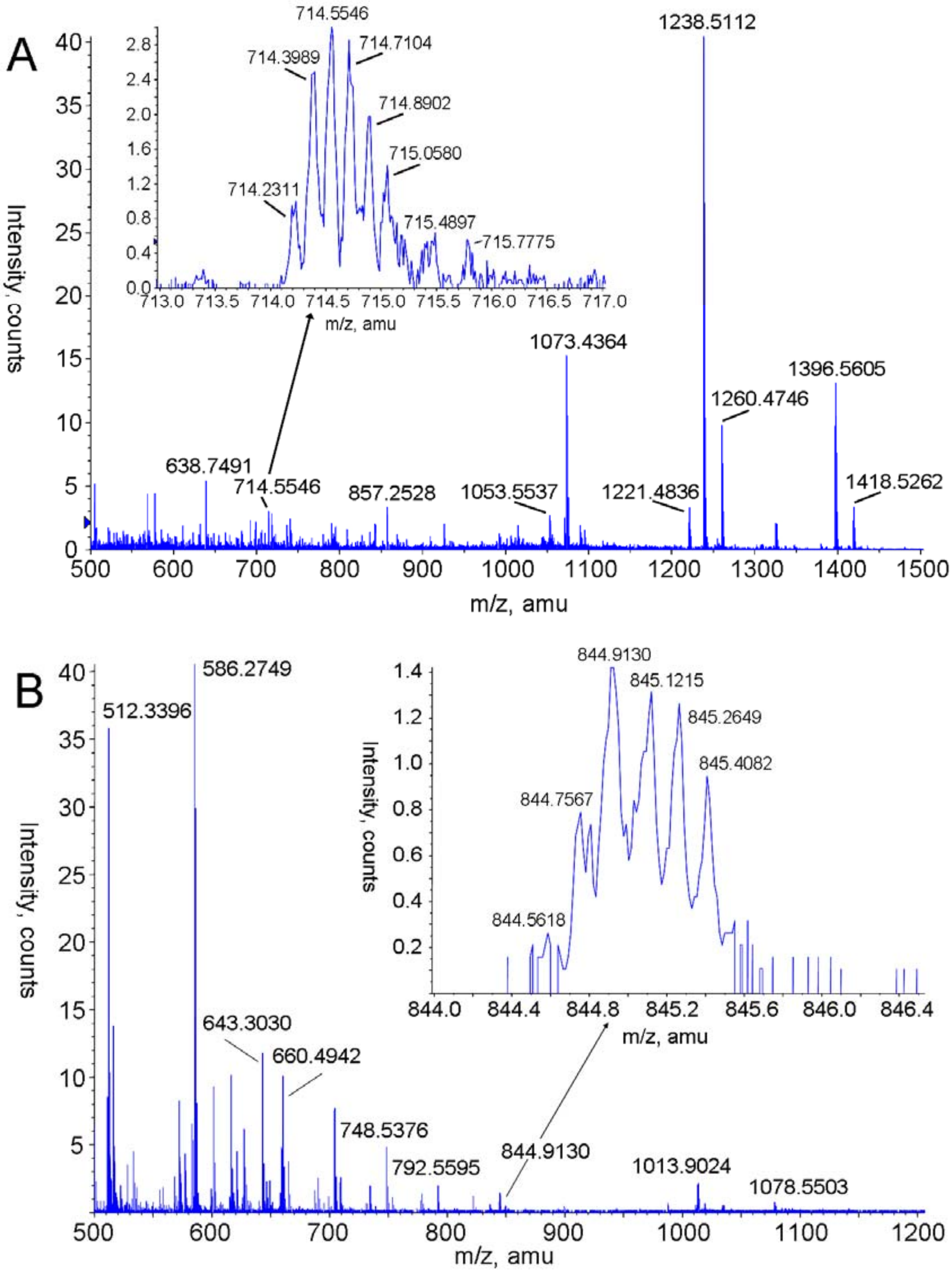


Figure 2.13. ^{13}**C isotope envelope and relative intensity of serum peptide peaks at m/z 714.2 and m/z 844.6.** These peptides had intensities of 2.8 and 1.4 respectively, relatively low compared to other molecular species that eluted at the same chromatographic time. The low Q1 resolution setting used stabilized the flight paths of molecules within ±2 m/z of the target peak selected. Thus, though the monoisotopic peaks for these species occurred at m/z 714.2 and m/z 844.6 in their +6 charge states, fragmentation was best achieved by targeting a peak within the isotope envelope having a higher m/z, in this case m/z 714.5 and m/z 844.9.

intensities (ion counts) of 2.8 and 1.4 in the +6 charge state when summed over 120 seconds (Figure 2.13).

MS/MS Analysis

Frozen, unacidified supernatant was thawed and an aliquot of 88% formic acid equal to the volume of sample was added. For each LC-MS/MS run, 0.5-2 µg of the acidified processed serum supernatant was hand injected onto the column to avoid sample loss in the injection loop. MS/MS fragmentation data were collected for m/z 70 to 2000 beginning at 17.92 min and ending at 19.92 min for the peak at m/z 844.6 and beginning at 17.69 min and ending at 19.69 min for the peak at m/z 714.2, these times marking the chromatographic window where the index peptide eluted as determined from the cLCMS run (Figure 2.14). Parent ion passage into the collision chamber was done with Q1 resolution set at "LOW" and the multi channel analyzer (MCA) was turned on. As part of this approach to obtain better sequence coverage of fragmentation patterns, 0.5 µg of sample was run seven times with collision energies (CE) of 15, 17, 20, 22, 25, 27, and 30 for the 714.2 m/z peak. These values correspond to the amount of the voltage drop (eV) that accelerates ions through the collision cell. Higher voltages produce greater acceleration and greater momentum at impact with inert gas (N_2) in the collision cell of the mass spectrometer. Higher values for the CE correspond to higher amounts of collision energy and smaller daughter ion fragments. The fragmentation of the 844.6 m/z peak was acquired using 1 µg of sample for each of 3 runs with a CE of 30, 35, and 37 and 2 µg of sample for each of 4 runs with a CE of 25, 27, 40 and 45.

Using the "Overlay" feature of the Analyst software the seven summed MS/MS spectra that were collected at the different collision energies were overlaid for each

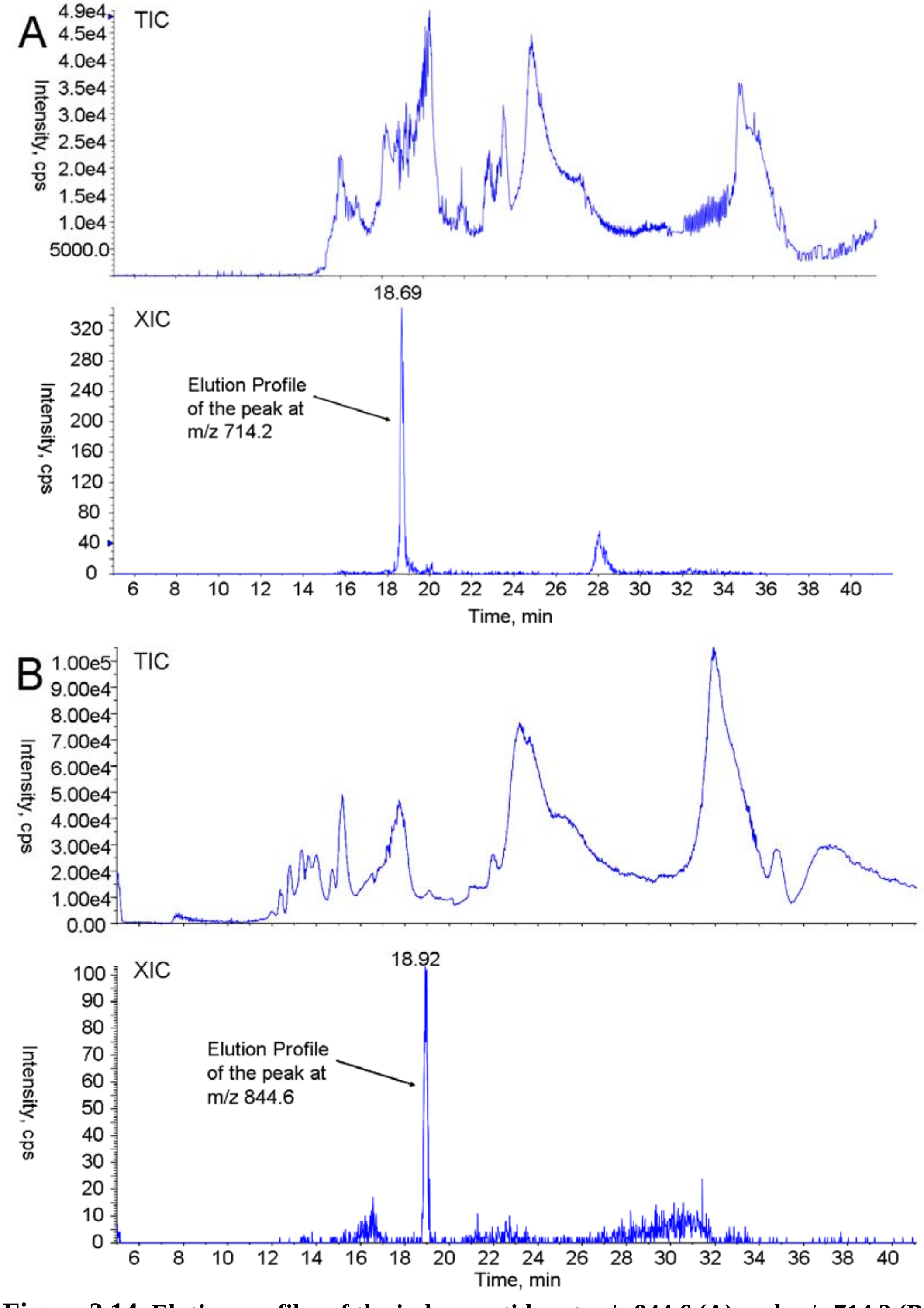


Figure 2.14. Elution profiles of the index peptides at m/z 844.6 (A) and m/z 714.2 (B). Total ion chromatogram (TIC) of 0.5 µg of total serum sample (top) and the extract ion chromatogram (XIC) of the range ± 0.75 m/z units from the peaks of interest (bottom) are shown. The peptides chosen for this study eluted at 18.92 ± 1.0 min (m/z 844.6) and 18.69 ± 1.0 min (m/z 714.2).

parent ion. The “Add Data” feature was then used to compile or combine data without averaging these seven MCA spectra together, giving a single MS/MS spectrum with fragmentation coverage over much of the sequence of the peptide (Figure 2.15). After the spectrum was smoothed, the threshold was set at 1.5 and the data was centroided. The centroided data threshold was set at 3.0 and the data list was exported to Excel.

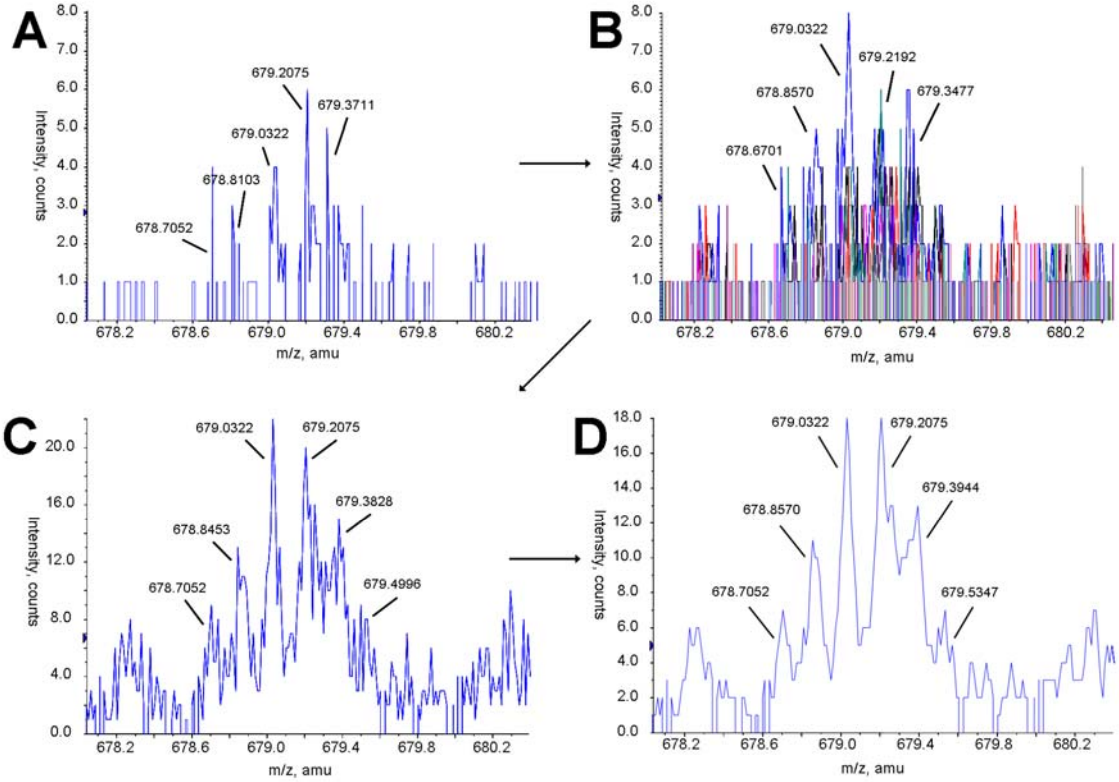
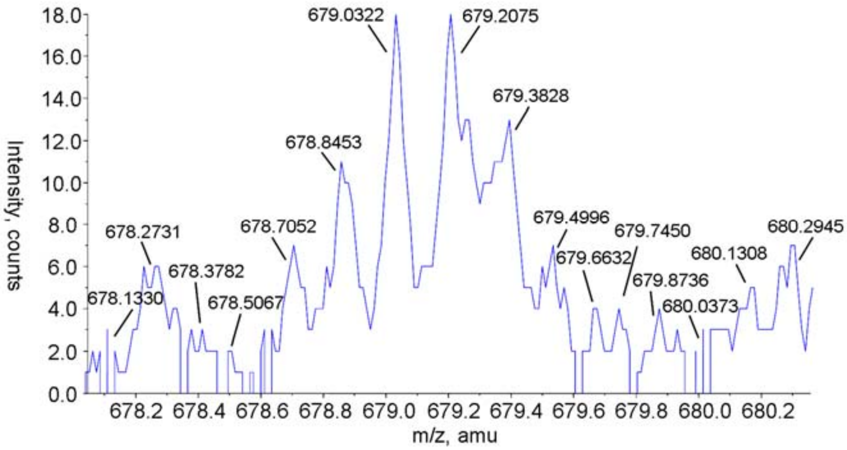


Figure 2.15 Addition and smoothing of 7 MCA MS/MS fragmentation spectra. Individual runs at single collision energies had relatively low levels of signal-to-noise (A). The spectra of several runs of the same specimen at different staged collision energies were overlaid (B) and then summed (C). The resultant compiled spectrum was then smoothed prior to sequence analysis (D). This procedure allowed for more comprehensive fragmentation coverage and increased signal without significantly increasing baseline noise.

The spectrum was visually inspected and compared to the exported data list to ensure the software had assigned charge states correctly (Figure 2.16). A previous study has shown that identification has been improved after deconvolution of the raw data. (Mujezinovic 2006) After any corrections, we transformed the data list using the formula: (+1 mass) = m/z value * (charge – (charge – 1)) so all species with charge states above +2 had a +1 mass listed. Peaks with undefined charges were considered to have +1 charge states. This corrected list was exported from Excel as a tab-delimited text file.



m/z	Assigned Charge	Corrected Charge
678.1330	4	1
678.2731	1	1
678.3782	1	1
678.5067	4	1
678.7052	3	6
678.8453	3	6
679.0322	3	6
679.2075	3	6
679.3828	3	6
679.4996	3	6
679.7450	3	6
679.8736	3	6
680.0373	3	6
680.1308	1	1
680.2945	2	2

Figure 2.16. Peak charge assignment correction. The data list for the compiled spectrum was exported to Excel and then visually checked to make sure the software had assigned charge states correctly. The software was usually accurate for lower charge states, but often had difficulty correctly assigning higher charge states. The +6 charged molecular species shown had been labeled as being a +3 species.

Regardless of the charge state of the parent ion, Mascot (www.matrixscience.com) only searches for matches to +1 and +2 fragments. Converting all fragments with charges of +3 and above to +1 allows Mascot to match those fragments. The text file was edited to have the following format: SEARCH=MIS; REPTYPE=Peptide; BEGIN IONS PEPMASS=844.6 (tab delimited data list (m/z intensity); END IONS. This text file was saved as a .tmp file, and this file was submitted to a Mascot MS/MS ions search. We searched the NCBInr database, limiting the search to mammalian sequences. The enzyme setting was set to ‘none’ with peptide and MS/MS tolerances of ± 1.0 Da, with a +6 peptide charge selected. The data was in Mascot generic format and the instrument used was designated as ESI-QUAD-TOF. Comparable amounts of serum were loaded and the peak of m/z 714.2 was targeted at CE 10, 17, and 25 to compare the intensity difference of spectra run with the MCA setting turned on and off.

MS/MS Sequence Analyses Performed by Standard IDA, by Customized Approach without MCA as Compared with the Full Customized Approach.

To better understand the utility of our method the same protein-depleted serum sample was submitted to the same cLCMS step followed by different MS/MS sequencing approaches.

In the first of these analyses, 0.5 µg of the protein-depleted sample was introduced onto the cLC column as described above. Peptide selection and sequencing were investigated using information-dependent acquisition (IDA). In this mode the instrument uses a one-second MS survey scan followed by 3, three second CID fragmentations with a rolling collision energy employed during the fragmentation period. The IDA mode is programmed to carry out fragmentation of the three most abundant peptides in each of these scans. The masses of the collision fragments were put on an exclusion list for 180 seconds, which exceeded the typical peak elution time of 60 seconds. The resulting spectra were submitted to Mascot for identification as outlined above with the charge setting set to look for matches to +1, +2, and +3 peptides. Peaks not considered by IDA but of greater abundance than each of the low abundance index peptides studied in their relevant mass spectra were also numbered as an indication of the IDA method sensitivity.

In order to carry out a more direct comparison of fragmentation methods, one of the 5 peptides identified when using IDA in the previous experiment was also submitted to the customized sequencing approach described above but with the MCA turned off as well as on. Analyses were carried out on 0.5 µg of the same protein-depleted serum sample, submitted first to cLCMS to locate the target peptide. The location of the target peptide was determined. The peak used represented a peptide in a +2 charge state having an m/z of 733.4. Then the peak was submitted to sequence analysis, collecting 120 MS/MS scans of the target peptide peak at a constant CE of 30, with the MCA both off and on. The fragmentation data was processed as described above and submitted to Mascot for possible peptide identification.

Results

Serum Proteomic cLC-MS

Protein-depleted serum was fractionated by gradient, capillary LC as described above and introduced into a time-of-flight mass detector by electrospray ionization. Approximately 4000-5000 distinct molecular ions were detectable using this approach. Based on low peak height, presence of a multiply charged state, being consistently found, and having amino acid sequence lengths of ~40-50 amino acids, two low abundance peptides were arbitrarily selected for further MS/MS analysis. The peptides chosen for this proof-of-principle study had masses of 5061.6 Da (m/z 844.6) and 4279.2 Da (m/z 714.2), with relative intensities (ion counts) less than 3 and 2 in the +6 charge state (Figure 2.13), which consistently eluted at 18.69 ± 1.0 min and 18.92 ± 1.0 min from the cLC-MS column using the gradient outlined above (Figure 2.14).

cLC-MS/MS Fragmentation

The amino acid sequences of these peptides were subsequently determined by cLC-MS/MS. Because the duty cycle of the instrument is such that one spectrum was taken every second, each two-minute elution window produced 120 total MS/MS spectra. With the Q1 resolution setting at ‘LOW’, the flight paths of a small range of m/z values

(±2 m/z) were stabilized through the quadrupole and subjected to collision cell

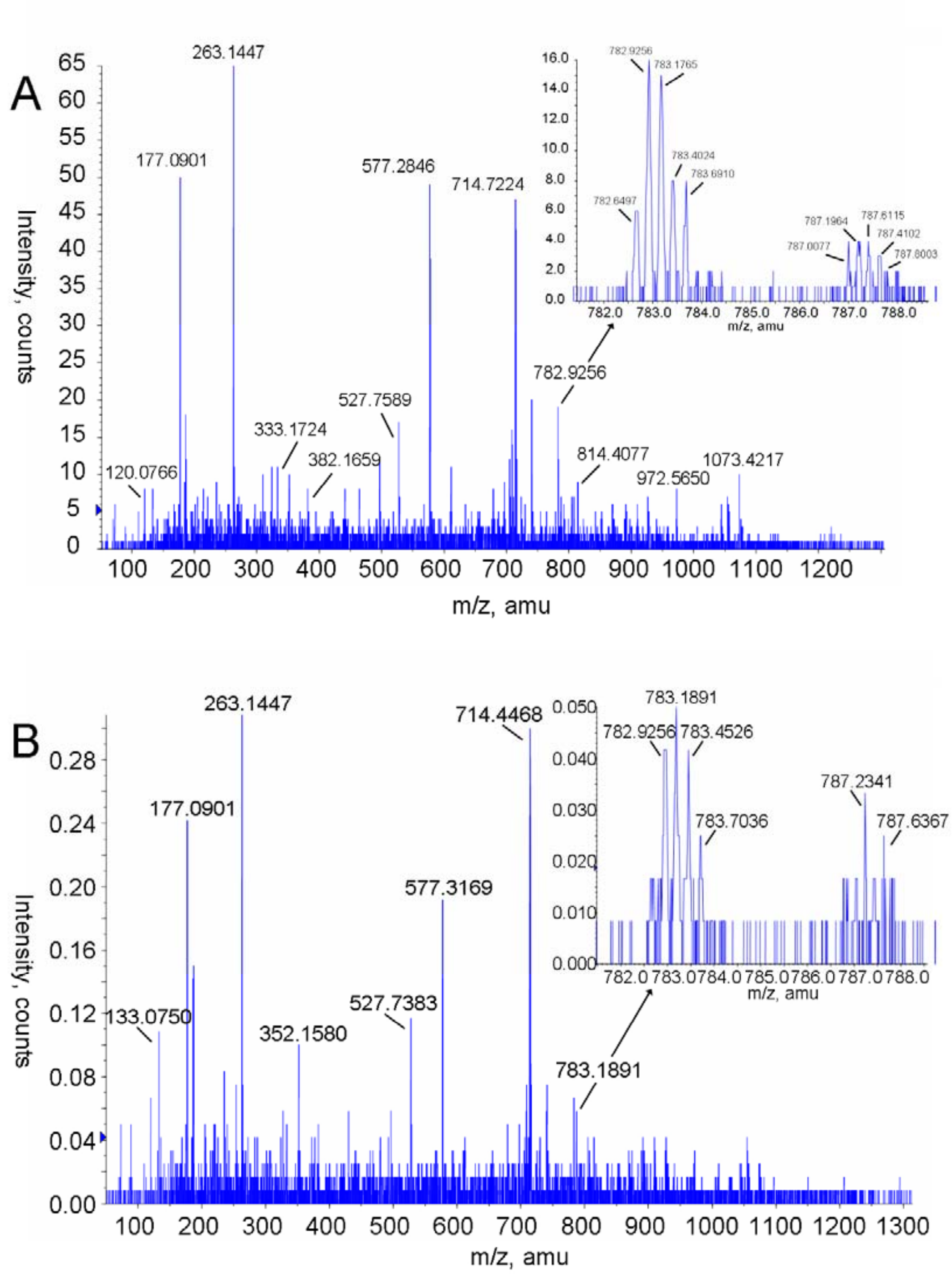


Figure 2.17. Comparison of fragmentation collected with the MCA feature turned on (A) and off (B). Without using the MCA feature, the maximum signal level reached 0.30 while the signal reached a maximum of 65 with the MCA turned on. Peak shape was marked improved, but signal to noise was only improved two to three fold.

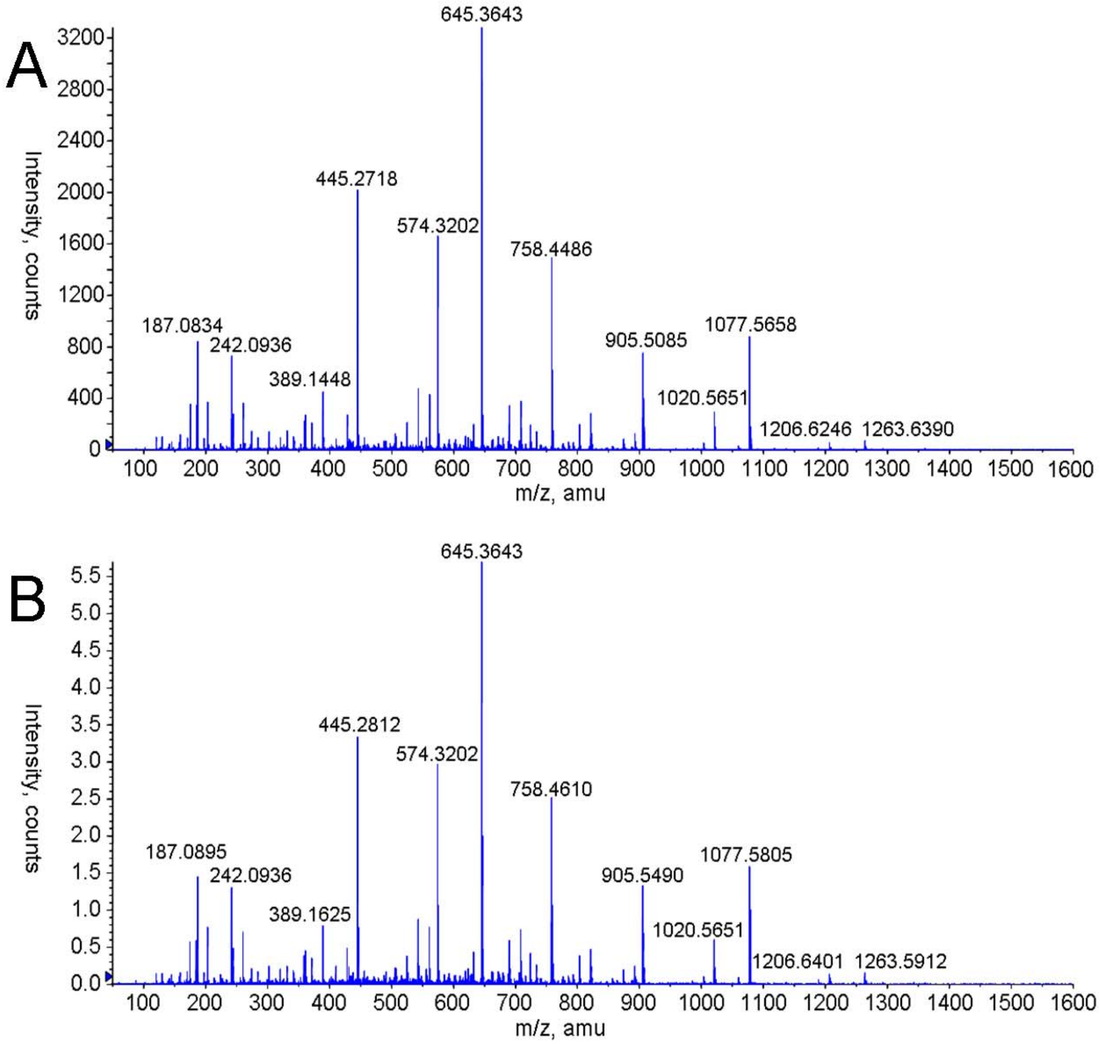


Figure 2.18. Fragmentation of the peptide, m/z 733.4 at CE 30 collected both with (A) and without (B) the MCA function turned on. This peptide was one of five identified using the IDA function for peptide sequencing. It was rerun using our approach for the sake of comparison. Although the MCA setting increased the intensity by over 500 times, both approaches also identified the peptide as the same specific fragment of fibrinopeptide A when submitted to Mascot.

fragmentation. Fragmentation of a large portion of the isotope envelope was best achieved by targeting a peak in the isotope window slightly above the monoisotopic peak, in this case 714.5 and 844.9. The MCA setting in the instrument method profile allowed for the summation of all 120 MS/MS spectra as they were taken. These MCA-summed spectra greatly increased MS/MS fragmentation intensities (Figures 2.17, 2.18). Several runs of the same specimen, each at different staged collision energies between 15 and 30 for peak 714.2 and between 25 and 45 for 844.6, were carried out. Higher collision energies produced smaller fragments and lower collision energies produced larger fragments, allowing for better coverage of the entire sequence (Figure 2.19). Summed spectra (120) from individual runs at each of 7 collision energies were then compiled to provide even further signal and signal-to-noise enhancement (Figure 2.16). This allowed for daughter ion fragment coverage of 33 of the 39 amino acids for the peptide of m/z 714.2 and of 42 out of 47 amino acids for the peptide with m/z of 844.6 peptide (Tables 2.4 and 2.5).

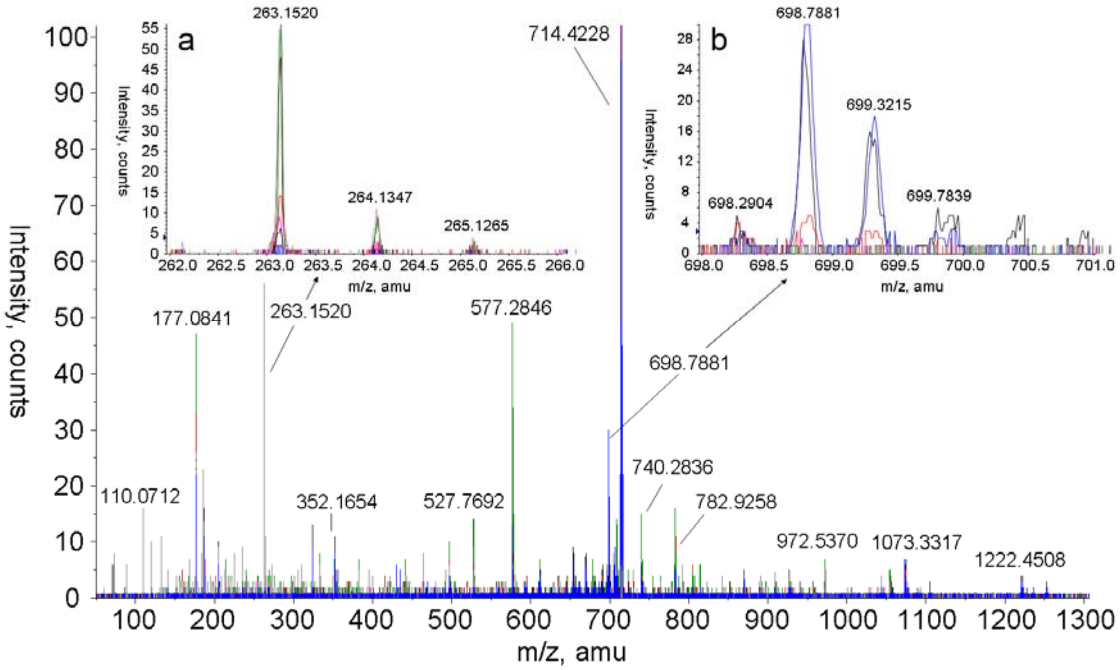


Figure 2.19. Peak fragmentation at different collision energies. The index peptide at m/z 714.2 was selected for fragmentation at seven different CEs. The seven spectra were overlaid: blue = CE 15, black = CE 17, red = CE 20, purple = CE 22, green = CE 25, maroon = CE 27, grey = CE 30. Higher CEs produced smaller fragments (a) while lower CEs produced larger fragments (b). The use of multiple collision energies allowed for more complete fragmentation of the parent ion.

MS/MS Data Analysis

Those fragmentation data representing ions with charges of +3 or greater were converted to the +1 ion mass using Excel as described above and compiled manually into a text file which was then submitted to a searchable sequence database, Mascot. The results of the submission are summarized in Tables 2.4 and 2.5. The amino acid sequences were successfully deduced by Mascot from the peptide fragmentation data. Specifically, the 714.2 peptide was identified as having a peptide sequence of **VSCKTSGYTFTEHGITWVRQAPGQGLECMGWISAHNGN** which is a peptide fragment of the immunoglobulin G heavy chain variable region. The 844.6 peptide was identified as having a amino acid sequence of **TFPGFFSPMLGEFVSETESRGSGSGIFTNTKESSSHHPGIAEFPSRG** which is a fibrin alpha isoform C-terminal fragment (Figure 2.20). The sequences were also submitted to a BLAST search on the National Center for Biotechnology Information site, which confirmed the initial identification (Data not shown). Standard IDA fragmentation did not result in selection of these peaks for MS/MS analysis and consequently did not provide their amino acid sequences. Moreover, a single spectrum taken at the apex of the elution profile for peak 714.2 showed 40 co-eluting peaks with intensities higher than it also not selected for IDA. Similarly, 23 coeluting peaks had intensities greater than that of peak 844.6 and were not selected by IDA. For these particular peptides, having the MCA off also resulted in inadequate peak height and data quality for sequence identification.

When 0.5 μg of the ACN-precipitated serum was subjected to fragmentation using IDA with a rolling collision energy, 488 total peptides were selected for fragmentation. However, only 5 of the 488 fragmentation studies resulted in protein identification when the spectra were submitted to Mascot. In order to provide a comparison of the instrument's standard IDA approach to peptide sequencing to our customized method, one of the peaks identified by IDA was further studied by our method. For this peptide two additional modes of operation were tried in an effort to define the benefit of using the MCA function. Using the peptide peak having a +2 charge state at m/z 733.4 for these studies, we collected 120 MS/MS scans of the peak at a constant CE of 30, with the MCA both on and off (Figure 2.21). The signal intensity of the fragmentation spectra collected with MCA on was over 500 times higher than that collected without MCA. Nevertheless, given the high abundance of the peptide, both fragmentation sets also resulted in spectra

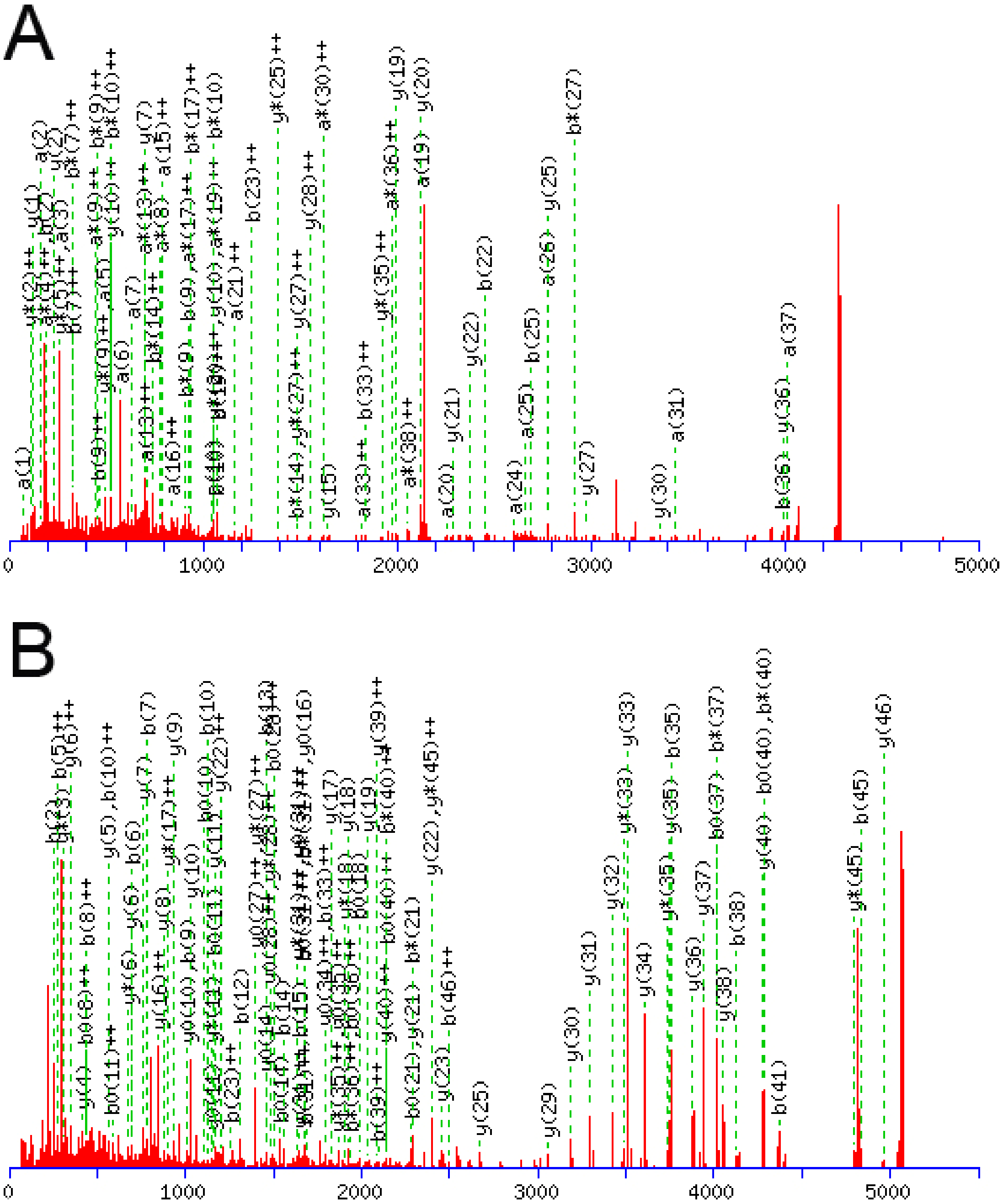


Figure 2.20. Fragmentation spectra of the index peptides 714.2 and 844.6 produced by submitting charge-corrected peak lists to Mascot. Green lines show fragmentation assignments. A and b ions are those that retained the charge on the N terminus, and y ions are those that retained the charge on the C terminus. Those peaks that represent fragments that retain two positive charges are denoted with a ++, those which have lost ammonia are denoted with an *, and those which have lost water with a º.

of sufficient coverage and intensity for Mascot identification, even with the single CE (Figure 2.18). This peptide was identified by all three approaches to be a fragment of fibrinopeptide A.

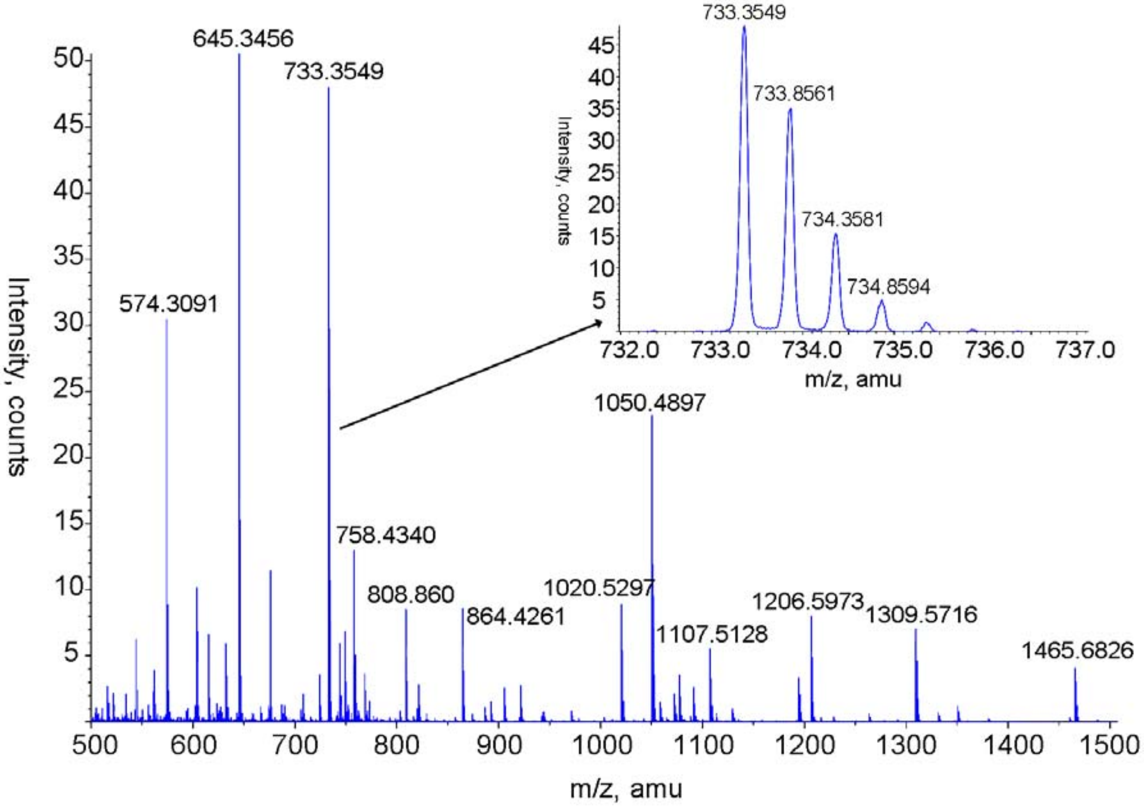


Figure 2.21. **^{13}C isotope envelope and relative intensity of the peptide at m/z 733.4** selected for method comparison. The peptide chosen for this study had a charge state of +2 and an intensity of 48, relatively high compared to other molecular species that eluted at the same time.

Discussion

The goal of this study was to demonstrate that effective serum proteomics could be carried out using capillary liquid chromatography, electrospray-ionization time-of-flight mass spectrometry and that the initial cLCMS step could be followed seamlessly with cLCMS/MS analyses to sequence and identify substantially larger peptides than had previously been reported using CID for fragmentation. This approach is especially attractive because the cLCMS/MS peptide sequencing uses the same column, gradient and MS instrumentation and does not require complete purification of the candidate, a digestion step, development of a second LC step, characterization of SELDI surface chemistry and potentially other steps that may be necessitated by using a second MS approach to sequence peptides of interest.

The use of cross-platform analysis adds complexity to some intact-protein-based, “top down” methods of biomarker discovery and identification. (Bogdanov 2005) Intact protein-based approaches use individual protein mass and elution characteristics to locate a single species and allow for evaluation of quantitative differences between comparison groups that might make it a useful biomarker. The larger size of proteins in an intact-protein-based approach, as well as the high potential for increased protein variability due to post-translational modifications, makes identification of serum proteins of interest by MS/MS most often prohibitively difficult. Additionally, both fragmentation and subsequent MS/MS amino acid sequence analysis become more complicated and less sensitive with increasing peptide length. (Loo 1988, Horn 2000) At the same time, it is becoming increasingly evident that peptide fragments of larger proteins may also prove to be important, perhaps even better, biomarkers. For example, the most widely used marker for prostate cancer is prostate specific antigen (PSA), which has a molecular weight of ~30,000. (Diamandis 2004, Qian 2006) Current PSA assays determine concentration of the intact molecule, but recent studies have suggested that specific fragments of PSA provide better sensitivity and specificity. (Petricoin 2006, Mikolajczyk 2004) This underscores the potential importance of a specific peptide approach in the search for new biomarkers.

#	a	a^{++}	a*	a^{*++}	b	b^{++}	b*	b^{*++}	Seq.	y	y^{++}	y*	y^{*++}	#
1	**72.0808**	36.5440			100.0757	50.5415			V					39
2	**159.1128**	80.0600			**187.1077**	94.0575			S	4181.9015	2091.4544	4164.8749	2082.9411	38
3	**262.1220**	131.5646			290.1169	145.5621			C	4094.8695	2047.9384	4077.8429	2039.4251	37
4	390.2170	195.6121	373.1904	**187.0988**	418.2119	209.6096	401.1853	201.0963	K	**3991.8603**	1996.4338	3974.8337	1987.9205	36
5	**491.2646**	246.1360	474.2381	237.6227	519.2595	260.1334	502.2330	251.6201	T	3863.7653	1932.3863	3846.7388	**1923.8730**	35
6	**578.2967**	289.6520	561.2701	281.1387	606.2916	303.6494	589.2650	295.1361	S	3762.7176	1881.8625	3745.6911	1873.3492	34
7	**635.3181**	318.1627	618.2916	309.6494	663.3130	**332.1602**	646.2865	**323.6469**	G	3675.6856	1838.3464	3658.6591	1829.8332	33
8	798.3815	399.6944	**781.3549**	391.1811	826.3764	413.6918	809.3498	405.1785	Y	3618.6641	1809.8357	3601.6376	1801.3224	32
9	899.4291	450.2182	882.4026	**441.7049**	**927.4240**	**464.2157**	**910.3975**	**455.7024**	T	3455.6008	1728.3040	3438.5743	1719.7908	31
10	**1046.4975**	523.7524	1029.4710	515.2391	**1074.4925**	537.7499	**1057.4659**	**529.2366**	F	**3354.5531**	1677.7802	3337.5266	1669.2669	30
11	1147.5452	574.2762	1130.5187	565.7630	1175.5401	588.2737	1158.5136	579.7604	T	3207.4847	1604.2460	3190.4582	1595.7327	29
12	1276.5878	638.7975	1259.5613	630.2843	1304.5827	652.7950	1287.5562	644.2817	E	3106.4370	**1553.7222**	3089.4105	1545.2089	28
13	1413.6467	**707.3270**	1396.6202	**698.8137**	1441.6416	721.3245	1424.6151	712.8112	H	**2977.3944**	**1489.2009**	2960.3679	**1480.6876**	27
14	1470.6682	735.8377	1453.6416	727.3245	1498.6631	749.8352	**1481.6366**	**741.3219**	G	2840.3355	1420.6714	2823.3090	1412.1581	26
15	1583.7523	**792.3798**	1566.7257	783.8665	1611.7472	806.3772	1594.7206	797.8639	I	**2783.3141**	1392.1607	2766.2875	**1383.6474**	25
16	1684.7999	**842.9036**	1667.7734	834.3903	1712.7948	856.9011	1695.7683	848.3878	T	2670.2300	1335.6186	2653.2035	1327.1054	24
17	1870.8792	935.9433	1853.8527	**927.4300**	1898.8742	949.9407	1881.8476	**941.4274**	W	2569.1823	1285.0948	2552.1558	1276.5815	23
18	1969.9477	985.4775	1952.9211	976.9642	1997.9426	999.4749	1980.9160	990.9617	V	**2383.1030**	1192.0551	2366.0765	1183.5419	22
19	**2126.0488**	1063.5280	2109.0222	**1055.0147**	2154.0437	**1077.5255**	2137.0171	1069.0122	R	**2284.0346**	1142.5209	2267.0081	1134.0077	21
20	**2254.1074**	1127.5573	2237.0808	1119.0440	2282.1023	1141.5548	2265.0757	1133.0415	Q	**2127.9335**	1064.4704	2110.9069	**1055.9571**	20
21	2325.1445	**1163.0759**	2308.1179	1154.5626	2353.1394	1177.0733	2336.1128	1168.5601	A	**1999.8749**	1000.4411	1982.8484	991.9278	19
22	2422.1972	1211.6023	2405.1707	1203.0890	**2450.1921**	1225.5997	2433.1656	1217.0864	P	1928.8378	964.9225	1911.8113	956.4093	18
23	2479.2187	1240.1130	2462.1921	1231.5997	2507.2136	**1254.1104**	2490.1871	1245.5972	G	1831.7850	916.3962	1814.7585	907.8829	17
24	**2607.2773**	1304.1423	2590.2507	1295.6290	2635.2722	1318.1397	2618.2456	1309.6265	Q	1774.7636	887.8854	1757.7370	879.3721	16
25	**2664.2987**	1332.6530	2647.2722	1324.1397	**2692.2937**	1346.6505	2675.2671	1338.1372	G	**1646.7050**	823.8561	1629.6784	815.3429	15
26	**2777.3828**	1389.1950	2760.3562	1380.6818	2805.3777	1403.1925	2788.3512	1394.6792	L	1589.6835	795.3454	1572.6570	786.8321	14
27	2906.4254	1453.7163	2889.3988	1445.2031	2934.4203	1467.7138	**2917.3938**	1459.2005	E	1476.5995	738.8034	1459.5729	730.2901	13
28	3066.4560	1533.7317	3049.4295	1525.2184	3094.4510	1547.7291	3077.4244	1539.2158	C	1347.5569	674.2821	1330.5303	665.7688	12
29	3197.4965	1599.2519	3180.4700	1590.7386	3225.4914	1613.2494	3208.4649	1604.7361	M	1187.5262	594.2667	1170.4997	585.7535	11
30	3254.5180	1627.7626	3237.4914	**1619.2494**	3282.5129	1641.7601	3265.4864	1633.2468	G	**1056.4857**	**528.7465**	1039.4592	520.2332	10
31	**3440.5973**	1720.8023	3423.5708	1712.2890	3468.5922	1734.7997	3451.5657	1726.2865	W	999.4643	500.2358	982.4377	**491.7225**	9
32	3553.6814	1777.3443	3536.6548	1768.8310	3581.6763	1791.3418	3564.6497	1782.8285	I	813.3850	407.1961	796.3584	398.6828	8
33	3640.7134	**1820.8603**	3623.6869	1812.3471	3668.7083	**1834.8578**	3651.6818	1826.3445	S	**700.3009**	350.6541	683.2743	342.1408	7
34	3711.7505	1856.3789	3694.7240	1847.8656	3739.7454	1870.3764	3722.7189	1861.8631	A	613.2689	307.1381	596.2423	298.6248	6
35	3848.8094	1924.9084	3831.7829	1916.3951	3876.8043	1938.9058	3859.7778	1930.3925	H	542.2318	271.6195	525.2052	**263.1062**	5
36	3962.8524	1981.9298	3945.8258	**1973.4165**	**3990.8473**	1995.9273	3973.8207	1987.4140	N	405.1728	203.0901	388.1463	194.5768	4
37	**4019.8738**	2010.4405	4002.8473	2001.9273	4047.8687	2024.4380	4030.8422	2015.9247	G	291.1299	146.0686	274.1034	137.5553	3
38	4133.9167	2067.4620	4116.8902	**2058.9487**	4161.9117	2081.4595	4144.8851	2072.9462	N	**234.1084**	117.5579	217.0819	**109.0446**	2
39									T	**120.0655**	60.5364			1

Table 2.4. Mascot Search Results for 714.2 peak. The corrected data list was submitted to a MASCOT database searching the NCBInr database. The amino acid sequence was successfully identified as a fibrinogen, or fibrin alpha isoform C-terminal fragment. The 71 peptide fragments matched (in red) represent a sequence coverage of 33 of the 39 amino acids in the peptide. Fragment ions in the a and b columns are those with the charge retained on the N terminus with the protein backbone fragmented at the C_{α}-CO bond or the peptide amide bond respectively. Fragment ions in the y column are those with the charge retained on the C terminus with the protein backbone fragmented at the peptide amide bond. In addition, peaks are seen for ions that retain two positive charges denoted a^{++}, b^{++}, and y^{++}, and those which have lost ammonia (-17) denoted a*, b*, and y*.

#	b	b^{++}	b*	b^{*++}	b^{0}	b^{0++}	Seq.	y	y^{++}	y*	y^{*++}	y^{0}	y^{0++}	#
1	102.0550	51.5311			84.0444	42.5258	T							47
2	**249.1234**	125.0653			231.1128	116.0600	F	**4961.3111**	2481.1592	4944.2845	2472.6459	4943.3005	2472.1539	46
3	346.1761	173.5917			328.1656	164.5864	P	4814.2427	2407.6250	**4797.2161**	**2399.1117**	4796.2321	2398.6197	45
4	403.1976	202.1024			385.1870	193.0971	G	4717.1899	2359.0986	4700.1633	2350.5853	4699.1793	2350.0933	44
5	550.2660	**275.6366**			532.2554	266.6314	F	4660.1684	2330.5879	4643.1419	2322.0746	4642.1579	2321.5826	43
6	**697.3344**	349.1708			679.3238	340.1656	F	4513.1000	2257.0536	4496.0735	2248.5404	4495.0895	2248.0484	42
7	**784.3664**	392.6869			766.3559	383.6816	S	4366.0316	2183.5194	4349.0051	2175.0062	4348.0210	2174.5142	41
8	881.4192	**441.2132**			863.4086	**432.2080**	P	**4278.9996**	**2140.0034**	4261.9730	2131.4902	4260.9890	2130.9981	40
9	**1012.4597**	506.7335			994.4491	497.7282	M	4181.9468	**2091.4770**	4164.9203	2082.9638	4163.9363	2082.4718	39
10	**1125.5437**	**563.2755**			**1107.5332**	554.2702	L	**4050.9063**	2025.9568	4033.8798	2017.4435	4032.8958	2016.9515	38
11	1182.5652	591.7862			**1164.5546**	**582.7810**	G	**3937.8223**	1969.4148	3920.7957	1960.9015	3919.8117	1960.4095	37
12	**1311.6078**	656.3075			1293.5972	647.3022	E	**3880.8008**	1940.9040	3863.7743	1932.3908	3862.7903	1931.8988	36
13	**1458.6762**	729.8417			1440.6656	720.8365	F	**3751.7582**	1876.3828	**3734.7317**	**1867.8695**	3733.7477	**1867.3775**	35
14	**1557.7446**	779.3759			**1539.7340**	770.3707	V	**3604.6898**	1802.8485	3587.6633	1794.3353	3586.6793	**1793.8433**	34
15	**1644.7766**	822.8920			1626.7661	813.8867	S	**3505.6214**	1753.3143	**3488.5949**	1744.8011	3487.6108	1744.3091	33
16	1773.8192	887.4133			1755.8087	878.4080	E	**3418.5894**	1709.7983	3401.5628	1701.2851	3400.5788	1700.7930	32
17	1874.8669	937.9371			1856.8563	928.9318	T	**3289.5468**	**1645.2770**	3272.5202	**1636.7638**	3271.5362	**1636.2718**	31
18	2003.9095	1002.4584			**1985.8989**	993.4531	E	**3188.4991**	1594.7532	3171.4726	1586.2399	3170.4885	1585.7479	30
19	2090.9415	1045.9744			2072.9310	1036.9691	S	**3059.4565**	1530.2319	3042.4300	1521.7186	3041.4460	1521.2266	29
20	2247.0426	1124.0250	2230.0161	1115.5117	2229.0321	1115.0197	R	2972.4245	1486.7159	2955.3979	**1478.2026**	2954.4139	**1477.7106**	28
21	2304.0641	1152.5357	**2287.0375**	1144.0224	**2286.0535**	1143.5304	G	2816.3234	1408.6653	2799.2968	**1400.1521**	2798.3128	**1399.6600**	27
22	2391.0961	1196.0517	2374.0696	1187.5384	2373.0856	1187.0464	S	2759.3019	1380.1546	2742.2754	1371.6413	2741.2914	1371.1493	26
23	2520.1387	**1260.5730**	2503.1122	1252.0597	2502.1282	1251.5677	E	**2672.2699**	1336.6386	2655.2433	1328.1253	2654.2593	1327.6333	25
24	2607.1707	1304.0890	2590.1442	1295.5757	2589.1602	1295.0837	S	2543.2273	1272.1173	2526.2008	1263.6040	2525.2167	1263.1120	24
25	2664.1922	1332.5997	2647.1657	1324.0865	2646.1816	1323.5945	G	**2456.1953**	1228.6013	2439.1687	1220.0880	2438.1847	1219.5960	23
26	2777.2763	1389.1418	2760.2497	1380.6285	2759.2657	1380.1365	I	**2399.1738**	**1200.0905**	2382.1473	1191.5773	2381.1632	1191.0853	22
27	2924.3447	1462.6760	2907.3181	1454.1627	2906.3341	1453.6707	F	**2286.0898**	1143.5485	2269.0632	1135.0352	2268.0792	1134.5432	21
28	3025.3924	1513.1998	3008.3658	1504.6865	3007.3818	**1504.1945**	T	2139.0213	1070.0143	2121.9948	1061.5010	2121.0108	1061.0090	20
29	3139.4353	1570.2213	3122.4087	1561.7080	3121.4247	1561.2160	N	**2037.9737**	1019.4905	2020.9471	1010.9772	2019.9631	1010.4852	19
30	3240.4830	1620.7451	3223.4564	1612.2318	3222.4724	1611.7398	T	**1923.9307**	962.4690	**1906.9042**	953.9557	1905.9202	953.4637	18
31	3368.5779	**1684.7926**	3351.5514	**1676.2793**	3350.5674	**1675.7873**	K	**1822.8831**	911.9452	1805.8565	**903.4319**	1804.8725	902.9399	17
32	3497.6205	1749.3139	3480.5940	1740.8006	3479.6100	1740.3086	E	1694.7881	**847.8977**	1677.7615	839.3844	**1676.7775**	838.8924	16
33	3584.6525	**1792.8299**	3567.6260	1784.3166	3566.6420	1783.8246	S	1565.7455	783.3764	1548.7190	774.8631	1547.7349	774.3711	15
34	3671.6846	1836.3459	3654.6580	1827.8327	3653.6740	1827.3406	S	1478.7135	739.8604	1461.6869	731.3471	**1460.7029**	730.8551	14
35	**3758.7166**	1879.8619	3741.6901	1871.3487	3740.7060	1870.8567	S	1391.6814	696.3444	1374.6549	687.8311	1373.6709	687.3391	13
36	3895.7755	1948.3914	3878.7490	**1939.8781**	3877.7650	**1939.3861**	H	1304.6494	652.8283	1287.6229	644.3151	1286.6388	643.8231	12
37	4032.8344	2016.9209	**4015.8079**	2008.4076	**4014.8239**	2007.9156	H	**1167.5905**	584.2989	**1150.5640**	575.7856	**1149.5799**	575.2936	11
38	**4129.8872**	2065.4472	4112.8606	2056.9340	4111.8766	2056.4419	P	**1030.5316**	515.7694	1013.5050	507.2562	**1012.5210**	506.7642	10
39	4186.9086	**2093.9580**	4169.8821	2085.4447	4168.8981	2084.9527	G	**933.4788**	467.2431	916.4523	458.7298	915.4683	458.2378	9
40	4299.9927	2150.5000	**4282.9662**	**2141.9867**	**4281.9821**	**2141.4947**	I	**876.4574**	438.7323	859.4308	430.2190	858.4468	429.7270	8
41	**4371.0298**	2186.0185	4354.0033	2177.5053	4353.0193	2177.0133	A	**763.3733**	382.1903	746.3468	373.6770	745.3627	373.1850	7
42	4500.0724	2250.5398	4483.0459	2242.0266	4482.0618	2241.5346	E	**692.3362**	**346.6717**	**675.3097**	338.1585	674.3256	337.6665	6
43	4647.1408	2324.0740	4630.1143	2315.5608	4629.1303	2315.0688	F	**563.2936**	282.1504	546.2671	273.6372	545.2830	273.1452	5
44	4744.1936	2372.6004	4727.1670	2364.0872	4726.1830	2363.5951	P	**416.2252**	208.6162	399.1987	200.1030	398.2146	199.6110	4
45	**4831.2256**	2416.1164	4814.1991	2407.6032	4813.2150	2407.1112	S	319.1724	160.0899	**302.1459**	151.5766	301.1619	151.0846	3
46	4987.3267	**2494.1670**	4970.3002	2485.6537	4969.3162	2485.1617	R	232.1404	116.5738	215.1139	108.0606			2
47							G	76.0393	38.5233					1

Table 2.5. Mascot Search Results for 844.6 m/z peak. The corrected data list was submitted to a MASCOT database searching the NCBInr database. The amino acid sequence was successfully identified as a fibrinogen, or fibrin alpha isoform C-terminal fragment. The 95 peptide fragments matched (in red) represent a sequence coverage of 42 of the 47 amino acids in the peptide. Fragment ions in the b column are those with the charge retained on the N terminus with the protein backbone fragmented at the peptide amide bond. Fragment ions in the y column are those with the charge retained on the C terminus with the protein backbone fragmented at the peptide amide bond. In addition, peaks are seen for ions that retain two positive charges denoted b^{++}, and y^{++}, those which have lost ammonia (-17) denoted b*, and y*, and those which have lost water (-18) denoted b^{o}, and y^{o}.

In contrast to IDA, cLCMS/MS fragmentation of a specific peptide can be accomplished with the MCA setting turned either off or on. With the MCA setting turned off, the method can be set up to take the periodic MS survey scans of all ions in the sample between MS/MS fragmentation scans so as to track the both progress of the run and the elution of specific peptides. Even though the duty cycle allows for one spectrum to be taken every second, the software only allows for consideration of the average of all the cLCMS/MS spectra taken within a selected time window. Summation of spectra is not possible with data collected using this method. Signal-to-noise ratios of individual cLCMS/MS spectra of low abundance peptides are low, and the signal-to-noise of averaged spectra are even lower. In contrast, with the MCA setting turned on, there is no option to use MS survey scans, and so run progress and elution times cannot be tracked simultaneously, but spectra can be summed as they are taken within a specified time window. This increases signal-to-noise ratios even more than summation of spectra. The cLCMS/MS runs targeting a specific peptide can only use one CE setting per run, regardless of whether the MCA setting is turned on or off. However, multiple cLCMS/MS runs covering a range of CE values appropriate to the size of the peptide being targeted (determined empirically) can be carried out and compiled which allowed for both good signal-to-noise ratios of daughter ions as well as good fragmentation coverage over the length of the peptides (Figure 2.20). The use of the MCA was critical. When the MCA was off, the two index peptides considered were not identified by their fragmentation data. The increased intensity provided by the summed spectra made the spectra much easier to interpret. However, we recognize that not every low abundance peptide will be sequenced with this customized approach despite improved fragmentation coverage and peak intensities.

A more direct comparison of IDA to our method was made for one of the peptides identified by IDA. The peptide evaluated had a +2 charge at m/z 733.4 (Figure 2.21). Using our approach, we collected 120 MS/MS scans of this peak at a constant CE of 30, with the MCA both on and off. Although the signal intensity of the fragmentation spectra collected with MCA was over 500 times higher than that collected without MCA, both spectra produced fragmentation data of sufficient quality for identification, even with a single CE (Figure 2.18). Both of the customized approaches allowed for Mascot identification of this peptide as the same specific fragment of fibrinopeptide A as found using IDA.

Large peptides frequently appear as multiply charged species in the cLCMS spectra. Two complications arise from this. First, while Analyst and other software programs usually assign charge states of +1 and +2 correctly, peptide fragments with a charge of +3 or higher are often misclassified. This necessitated a manual comparison of the data list generated by many MS computer software analyses to the actual spectra to correct charge assignment mistakes. Second, Mascot considers only +1 and +2 daughter fragment ions when comparing experimental data to database protein sequences. For this reason, we manually converted all daughter fragment ions with charge states greater than +2 to their respective +1 m/z values before submitting the converted data list to Mascot. The MOWSE scoring algorithm used by Mascot to determine sequence identification confidence limits takes into account all possible theoretical fragments. This score is calculated as the logarithm ($-10 \times \log(p)$) of the probability (p) that the sequence ID that was made by comparing the fragmentation spectra to the database is a chance event (www.matrixscience.com). Despite fragments accounting for 42 of 47 amino acids for peak 844.6, this represented only 95 fragments out of a theoretical 508 possible and the Mascot search returned a Mowse score of 24. For this particular search, a score of > 23 indicated peptides with significant homology, while individual ions required scores > 63 to indicate identity or extensive homology ($p<0.05$). While a score of 24 suggested some homology, such a score would not, in and of itself, provide much confidence in the sequence matched to the spectra, but a protein BLAST search of the amino acid sequence deduced by MASCOT provided an E value of $5x10^{-18}$, indicating a high degree of certainty in peptide assignment.

In summary, the approach described here provides a dramatic enhancement in the peptide sequencing/protein identification capabilities of cLCMS plus cLCMS/MS serum proteomics. The possible application of this approach to serum biomarker discovery by making direct comparisons between sera of individuals with a disease to the sera of those who do not will require additional work. In particular, making serum proteomics quantitative will be important. Quantitation by means of MS is challenging and continues to be a major concern in the analysis of complex specimens, especially when coupled to specimen preprocessing, chromatographic inconsistencies, variable ionization efficiencies and fluctuations in instrument performance. Previous serum proteomic studies demonstrate these difficulties, but to date there have been no effective remedies for such potential problems. Clearly, future work will need to address these issues. (Mikolajczyk 2004) Due to the empirical nature of this method, it is not likely that this method would be easily automatable. As recently stated by Mujezinovic *et al.* "…no automated procedure will match the performance of the experienced eye and the intuition of an MS specialist in the foreseeable future." (Baggerly 2004) Hence, more sophisticated data analytical techniques represent another need for successful biomarker discovery.

In conclusion, the successful observation and identification of two index peptides of 39 and 47 amino acids from a single serum sample, using combined cLCMS and cLCMS/MS methods as part of a single platform exceed previously published size limits for LC-MS/MS and suggest that these techniques can be useful in clinical biomarker discovery.

Chapter 3 – Proteomic Identification of Serum Peptides Predicting Subsequent Spontaneous Preterm Birth

Abstract

Background

Current interventions to prevent spontaneous preterm birth (SPTB) are most efficacious if administered prior to advanced active labor. We used a proteomic approach to identify markers of SPTB in the serum of women at 24 and 28 weeks gestation prior to the onset of symptoms.

Methods

Serum proteomics was applied to sera from 80 pregnant women sampled at 23.4 weeks gestation (range 22-24 wks) and an additional 80 pregnant women sampled at 27.6 weeks (range 26-29 wks). Half of these women at each time had uncomplicated pregnancies with term deliveries and half had pregnancies resulting in SPTB. Additional analytes had been previous assayed in these specimens.

Results

Three specific peptides arising from inter-alpha-trypsin inhibitor heavy chain 4 protein were reduced significantly in the serum of women with later SPTB. The differences were observed for women at both 24 and 28 weeks gestation, approximately 8 and 4 weeks prior to SPTB. Levels of these peptides were lower the more imminent the SPTB. Likewise their positive predictive values were higher at 28 weeks. The peptide mass 2027 was the most discriminating with a sensitivity of 65% with a specificity of 82.5%, OR=8.8, CI: 3.1-24.8 at 28 weeks. The 3 current peptide biomarkers when combined

with 6 previously studied candidate biomarkers provided a specificity of 86.5% and a sensitivity of 80.5% at 28 wks.

Conclusion

Three novel serum markers of SPTB have been identified using a serum proteomic approach. Using these markers and 6 previously considered markers, women at greater risk of SPTB can be identified weeks prior to SPTB allowing for medical intervention.

Introduction

Spontaneous preterm birth (SPTB) is the leading cause of perinatal morbidity and mortality in the United States. (Rush 1976, Goldengerg 1998) Despite the magnitude of the problem and the substantial research efforts of many investigators, completely efficacious therapies for the treatment or prevention of SPTB have yet to be developed. Indeed, the rate of SPTB has not changed in decades. (Hoyert 2006) A major obstacle to the development of an effective treatment for preterm labor is a limited understanding of the molecular events required to initiate and maintain term and preterm labor.

Several proteins present in maternal serum or cervical secretions have been proposed as markers that may predict SPTB. Goldenberg and associates, in perhaps the most comprehensive study to date, evaluated a screening test consisting of three serum markers (CRH, AFP, alkaline phosphatase) and two cervical secretion markers (fetal fibronectin and ferritin). (Goldenberg 2001) This use of multiple markers increased sensitivity, specificity and the odds ratio of the predictive test. (Goldenberg 2001) However, none of the current SPTB markers alone or in combination provided adequate specificity or sensitivity to be used predictively.

Recent advances in technology allow for the evaluation of a large, unbiased portion of the complement of peptides and/or proteins present in maternal serum. Serum proteomic analysis, consisting of chromatographic separation followed by mass spectrometry to identify peptides and proteins by mass, can provide an extensive inventory of peptides and/or proteins present at any given time. Previous studies have attempted to use proteomic patterns to identify patients with early ovarian, breast and prostate cancers. (Buhimschi 2005) The use of proteomic analysis to identify phenotypic molecular characteristics of women who experience SPTB or infection has been attempted in amniotic fluid (Gravett 2004, Buhimschi 2005, Ruetschi 2005, Buhimschi 2007) and cervical secretions (Iams 1996, Ruetschi 2005, Gravett 2007, Pereira 2007), but serum proteomic analysis has not been reported.

SPTB is well suited for a proteomic approach given likely serologic changes that precede by weeks its clinical manifestations. We hypothesize that proteomic differences exist in maternal serum several weeks prior to the onset of clinical symptoms in women destined to develop SPTB. Our aim was to use serum proteomics to differentiate women having a later SPTB from those having term deliveries. Moreover, we hoped to identify all peptides that are found to be increased or decreased in the serum of women who go on to have a SPTB as compared to those who deliver at term.

Methods

Patient Population

This study represents a nested case-control study that used samples and data that were collected during the National Institute of Child Health and Human Development Maternal-Fetal Medicine Units Network Preterm Prediction Study (Merrell 2004). The Preterm Prediction Study, conducted between 1992 and 1994, was a multicenter observational investigation of 2929 symptom-free women evaluated longitudinally to determine risk factors for spontaneous preterm birth. Women were enrolled in this study without regard to specific risk factors for spontaneous preterm birth. Extensive information and biologic specimens were collected at each of 4 study visits, beginning at approximately 22 to 24 weeks' gestation and occurring at 2-week intervals. The overall study population and the methods used in the Preterm Prediction Study have been previously described in detail. Gestational age was based on the last menstrual period if the last menstrual period–derived gestational age was confirmed within 10 days by the earliest ultrasonographic evaluation. A spontaneous preterm birth was defined as a preterm birth < 35 weeks gestation occurring as the result of the spontaneous onset of labor or spontaneous ruptures of membranes.

Serum was collected at 24 and 28 weeks gestation and pregnancy outcomes were obtained. Participating women provided voluntary, informed consent. The original study protocols as well as these secondary analyses were approved by the representative institutional review boards. For this study, serum from 40 subjects who experienced a later SPTB and 40 subjects having uncomplicated pregnancies were obtained at 24 wks gestation (Visit 1) and submitted to proteomic analysis. Additionally, serum from 40 subjects who experienced a later SPTB and 40 subjects having uncomplicated pregnancies that ultimately delivered at term after spontaneous onset of labor was obtained at 28 wks (Visit 3) gestation and was likewise analyzed. Cases and controls were randomly selected to produce a representative group from among the cohort. Researchers were blinded to which group represented controls and which cases during the proteomic analysis and data evaluation.

Specimen Preparation

High molecular weight, and typically uninformative, proteins were removed by organic solvent precipitation as described in chapter 2.

Capillary Liquid Chromatography

Capillary liquid chromatography (cLC) was interfaced with a mass spectrometer, allowing for the continuous direct delivery of fractionated, protein-depleted serum to the mass detector. A complete description of this is provided in Chapter 2.

Electrospray-Ionization, Time-of-Flight Mass Spectrometry

Effluent from the cLC was directed into a QSTAR® Pulsar i quadrupole orthogonal time-of-flight mass spectrometer through an IonSpray source (Applied Biosystems). Mass spectra were collected every second for m/z 500 to 2500 from 5 to 55 min elution. Data collection and preliminary formatting were accomplished using Analyst QS® software with BioAnalyst add-ons (Applied Biosystems). Specimens from cases and controls were analyzed together in randomized order.

To reduce data file size, each mass chromatogram was divided into ten 2-minute elution intervals. One reference peak, observable in all specimens, near the center of each interval, and which did not demonstrate differences in abundance between the two groups was used to align time in that elution region. Of the ten elution intervals, the first to be analyzed (and the only one reported here) was the second 2 min window, chosen because more peptides were present. Systematic manual inspection of each case and control spectrum identified candidate species for further analysis. (Merrell 2008)

Evaluation of Candidate Molecular Species

Each species that appeared quantitatively different between groups was further evaluated. To facilitate determination of statistically significant quantitative differences, a second peak near the candidate peak but quantitatively comparable as determined by visual inspection of a sample of individual mass spectra in both groups was selected as an internal control. This reference peak was then used to normalize the candidate peak of interest, correcting for variability in specimen processing, specimen loading, ionization efficiency and instrument performance and allowing comparison across runs performed on different days. Thereafter, the molecular species were 'extracted' by the Analyst® software to determine a quantitative peak height of both candidate and reference ions in each specimen for all subjects.

Mass Spectral Data Analysis

The abundance (peak height) of each candidate and reference species was tabulated as was the calculated ratio of each candidate species abundance relative to reference species abundance within each patient. The log of that ratio was also determined because abundance varied substantially. The data were submitted to statistical analysis.

MS-MS Amino Acid Sequencing

Candidate species demonstrating statistically different abundances between cases and controls were further analyzed in an effort to chemically identify the candidate molecule. Frozen supernatant (from the protein reduction step) was thawed and hand injected for MS-MS analysis with a one sec TOF-MS scan being taken for m/z of 500 to 2500, followed by a 3 sec Positive Product Ion scan taken to the collision cell and the second MS sector to study the ion of interest. The selected ion was fragmented by N_2 gas collision and the daughter fragments 'read' in the second MS.

MS-MS fragmentation data were produced at several discrete collision energies, unique to the peptide's size, and all the fragment data summed. The spectrum were visually inspected and compared to the exported data list to insure the software had assigned charge states correctly. After any correction, the data list was transformed using the formula: +1 mass = m/z value * charge – (charge – 1) so that all species had a +1 mass listed. The corrected list was submitted to Mascot (www.matrixscience.com), a searchable MS database allowing protein/peptide identification. Amino acid sequences were also independently submitted to the short homologous or near homologous protein BLAST search available through the NCBI website as a confirmation.

Previous Biomarker Analysis

Plasma corticotrophin releasing factor, defensin, ferritin, lactoferrin, thrombin anti-thrombin complex and tumor necrosis factor α receptor type 1 assays have been previously described. (Goldenberg 2001)

Statistical Analysis

Data are expressed as means ± 1 standard error. Species that appeared to be quantitatively different were considered. Only three peaks were evaluated further. Comparisons of the abundance of a single species for the two study populations were carried out by the Wilcoxon rank sum test. Comparisons were carried out for each candidate at both 24 and 28 wks gestation. Twenty three additional serum markers were previously assayed in these samples and comparisons of the abundance of each individual marker were calculated using Wilcoxon rank sum test. Logistic regression analyses were performed for the three novel biomarkers. These three in combination with the best 6 of the previously tested markers were used for classification performance by means of receiver operator curves. For all statistical tests, nominal two-sided P-values are reported with statistical significance defined as a P-value < 0.05. SAS version 8.2 (SAS Institute, Cary, North Carolina) was used for these analyses.

Results

The demographics of the four groups (Case and Control at Visit 1 and Case and Control at Visit 2) are provided in Table 3.1.

	24 Weeks	24 Weeks	28 Weeks	28 Weeks
	Controls	Cases	Controls	Cases
	N=40	N=40	N=40	N=40
Maternal Age (year)	23.2 ± 0.83	23.6 ± 0.81	24.3 ± 0.9	24.2 ±0.94
GA at Delivery (wk)	38.9 ± 0.19	31.4 ± 0.44*	38.9 ± 0.18	32.3 ± 0.28*
Time from Visit 1 to Delivery (wk)	15.1 ± 0.18	7.8 ± 0.45*	15.1 ±0.18	8.6 ± 0.31*
Parity (% nulliparous)	30.0	32.5	30.0	32.5
Race (% African American)	75.0	70.0	77.5	67.5

Table 3.1. Demographics of four preterm birth groups (* p<0.001)

The individual mass spectra were averaged for the 2-minute window that was analyzed. Within these 10 cLC windows 4000-5000 unique ion peaks were detected. Our initial survey involved the second time window, selected because more peptides were observable.

After visual inspection, four molecular ions were further evaluated to determine if quantitative differences were significant. Of the 4 species considered further, 3 were found to be quantitatively significantly different (Table 3.2). Those species were: an ion at 676.66 m/z with a +3 charge corresponding to a neutral parent mass of 2026.98 Da; an ion at 856.85 m/z with a +5 charge corresponding to a neutral parent mass of 4279.25 Da; and an ion at 860.05 m/z with a +5 charge that corresponded to neutral parent masses of 4295.25 Da. The exact peak height of each of these species as well as a reference ion

24 Weeks Gestation

	Preterm Birth		Control		
Label	**Mean**	**Std Error**	**Mean**	**Std Error**	**P Value**
Peak at 677(/673) m/z	0.198	0.076	0.503	0.094	0.007
Peak at 857(/843) m/z	-0.137	0.086	0.284	0.092	0.002
Peak at 860(/843) m/z	-0.376	0.088	0.009	0.098	0.005

28 Weeks Gestation

	Preterm Birth		Control		
Label	**Mean**	**Std Error**	**Mean**	**Std Error**	**P Value**
Peak at 677(/673) m/z	-0.015	0.090	0.579	0.101	< 0.0001
Peak at 857(/843) m/z	-0.149	0.095	0.231	0.102	0.007
Peak at 860(/843) m/z	-0.204	0.088	0.201	0.095	0.002

Table 3.2. Ion abundance of the candidate peaks log ratioed to reference.

24 Weeks Gestation

	Preterm Birth		Control		
Label	**Mean**	**Std Error**	**Mean**	**Std Error**	**P Value**
Placental growth factor	446.82	45.99	596.69	63.36	0.05
Thrombin anti-thrombin	274.80	255.91	293.83	269.60	0.02

28 Weeks Gestation

	Preterm Birth		Control		
Label	**Mean**	**Std Error**	**Mean**	**Std Error**	**P Value**
Corticotropin releasing factor	0.3585	0.0189	0.2844	0.0122	0.0006
Defensin	612.0	106.6	427.4	92.1	0.039
Ferritin	18.97	3.37	10.00	1.48	0.043
Lactoferrin	245.0	42.0	484.6	100.7	0.046
Thrombin anti-thrombin	546.8	366.3	835.5	453.7	0.044
TNF receptor type 1	1114.1	119.0	880.4	33.7	0.018

Table 3.3. Comparison of relative abundance of other potential markers in the serum of cases and controls.

nearby was determined by the instrument. These same ion species were studied at both

24 and 28 wks gestation visits and were found to be significantly different between cases and controls at both gestational ages. Other potential markers with significant difference in abundance between cases and controls are listed in Table 3.3.

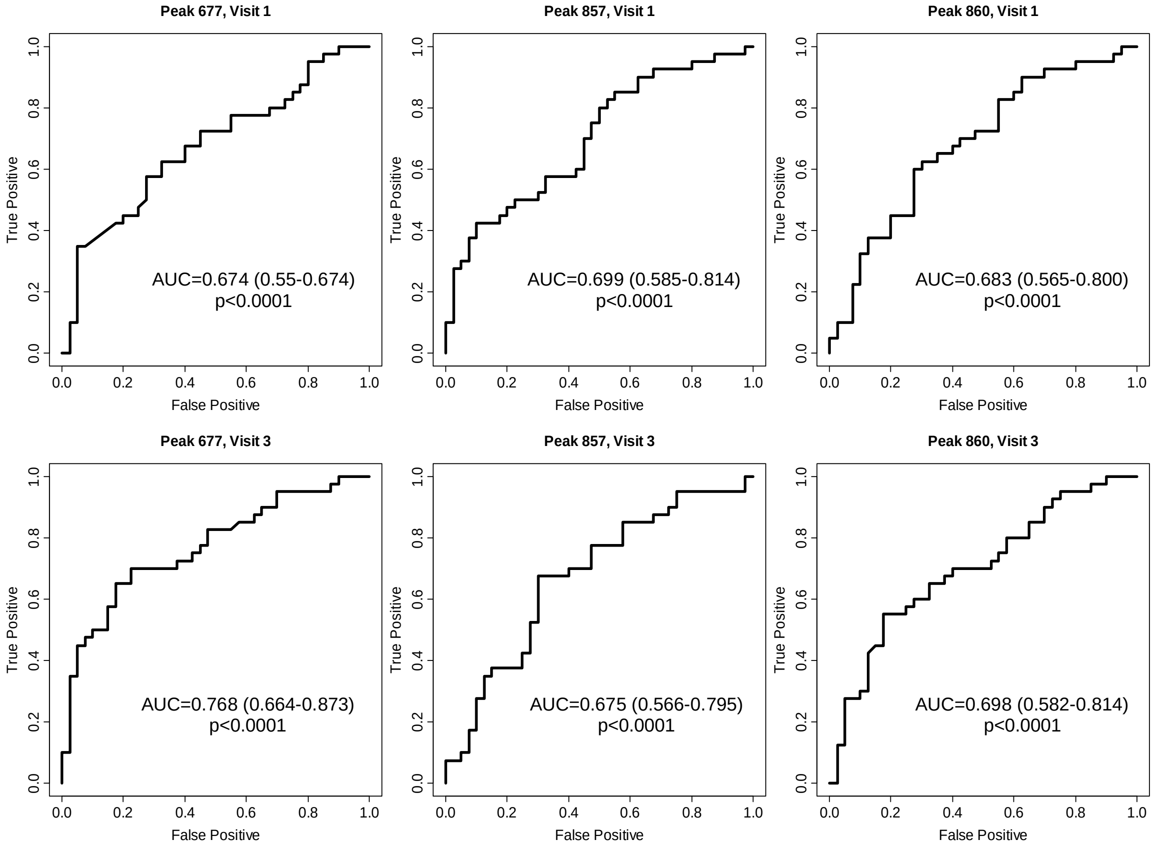


Figure 3.1 A. ROC curves demonstrating the predictive capability of 3 peptide markers to predict subsequent SPTB after sampling at 24 and 28 weeks. Area under the curve and 95% confidence intervals are also included for each marker at each visit.

Predictive value of the markers

The sensitivity of each of the three biomarkers improved generally from 24 to 28 weeks (At 24 weeks: 677, sensitivity=35.0%, specificity=92.5%, OR=6.64, CI 1.7-25.5; 857, sensitivity=45.0%, specificity=82.5%, OR=3.86, CI 1.4-10.8; 860,

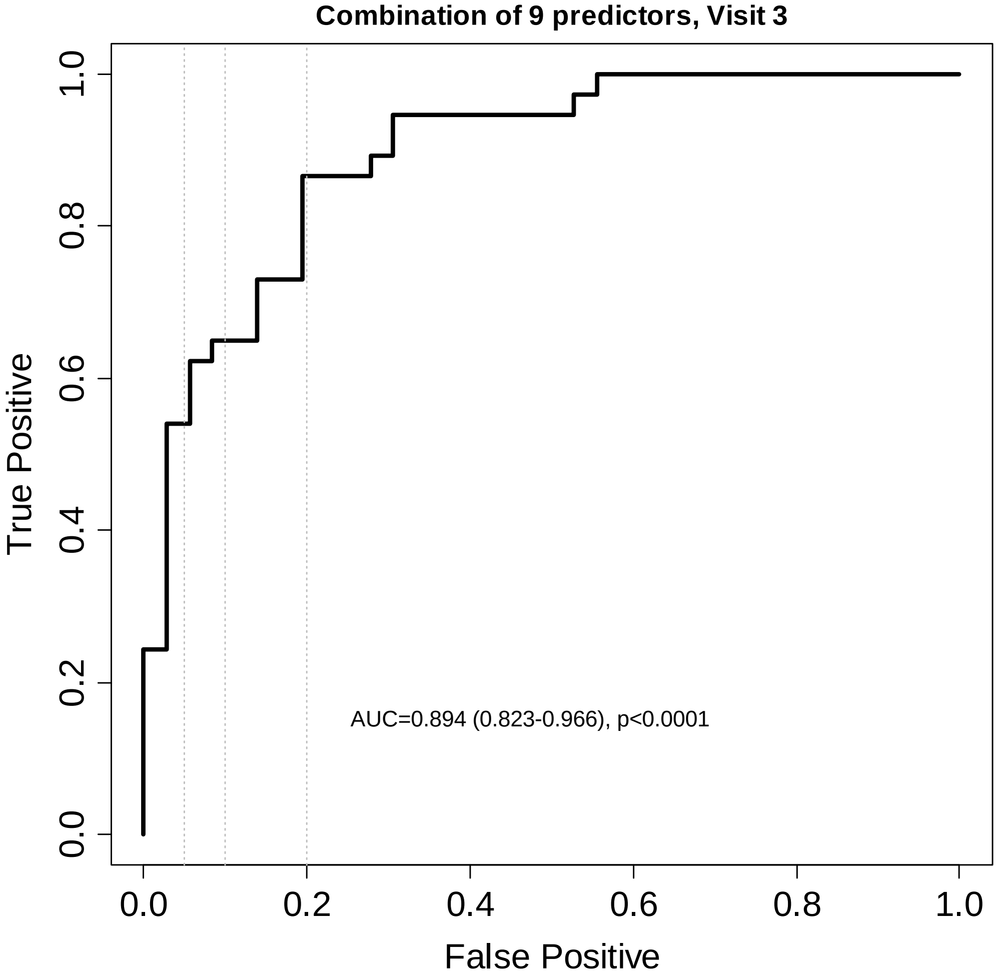


Figure 3.1 B. ROC curve demonstrating the predictive capability of the combination of 9 predictors including peak 677, peak 857, peak 860, corticotrophin releasing factor, defensin, ferritin, lactoferrin, thrombin antithrombin complex, and tumor necrosis factor – receptor type 1 to predict subsequent SPTB after sampling at 28 weeks. Area under the curve and 95% confidence intervals are also reported.

sensitivity=45.0%, specificity=80.0%, OR=3.27, CI 1.2-8.8. At 28 weeks: 677, sensitivity=65.0%, specificity=82.5%, OR=8.76, CI 3.1-24.8; 857, sensitivity=37.5%, specificity=80.0%, OR=2.4, CI 0.9-6.6; 860, sensitivity=55.0%, specificity=80.0%, OR=4.89, CI 1.8-13.2). The biomarker at m/z 676.7 was the best single predictor of SPTB at 24 or 28 wks pregnancy. (Figure 3.1 A) Combination of the 3 markers did not change the sensitivity and specificity but the inclusion of the 6 best additional markers improved the sensitivity to 86.5% with a specificity of 80.5% at 28 weeks (Figure 3.1 B).

Peptide Identification

Sequencing by means of a tandem MS/MS with intervening fragmentation allowed for the complete amino acid sequence to be determined by amino acid homology to known peptide or protein sequences. The amino acid sequences are provided in Table 3.4. The peaks initially assessed represented a +3 charge state for the species at 677 and a +5 charge state for the species at 857 and 860. Molecular ions representing additional charge states (+2 for 677, both +6 and +7 for 857 and 860) were also observed and were also quantitatively significantly reduced in the women with later SPTB (Data not shown). When a BLAST search of the individual amino acid sequences was performed using the National Center for Biotechnology Information website, all three peptides were found to be derived from one region of inter-alpha-trypsin inhibitor heavy chain 4 (ITIH4), the common parent protein.

m/z	Charge	MW	Amino Acid Sequence
677	+3	2026.98	qlglpgppdvpdhaayhpf
857	+5	4279.25	nvhsagaagsrmnfrpgvlssrqlglpgppdvpdhaayhpf
860	+5	4295.25	nvhsagaagsrm(O)nfrpgvlssrqlglpgppdvpdhaayhpf

Table 3.4. Amino acid sequences for the 3 biomarker peptides

Relationship of Biomarker Abundance as a Function of Time to Delivery

Women in the case group at 24 wks were on average 7.8±0.45 wks away from their actual preterm delivery, the mean gestational age at delivery being 31.4±0.44 weeks. Women in the case group at 28 wks were on average 4.7±0.32 wks removed from their

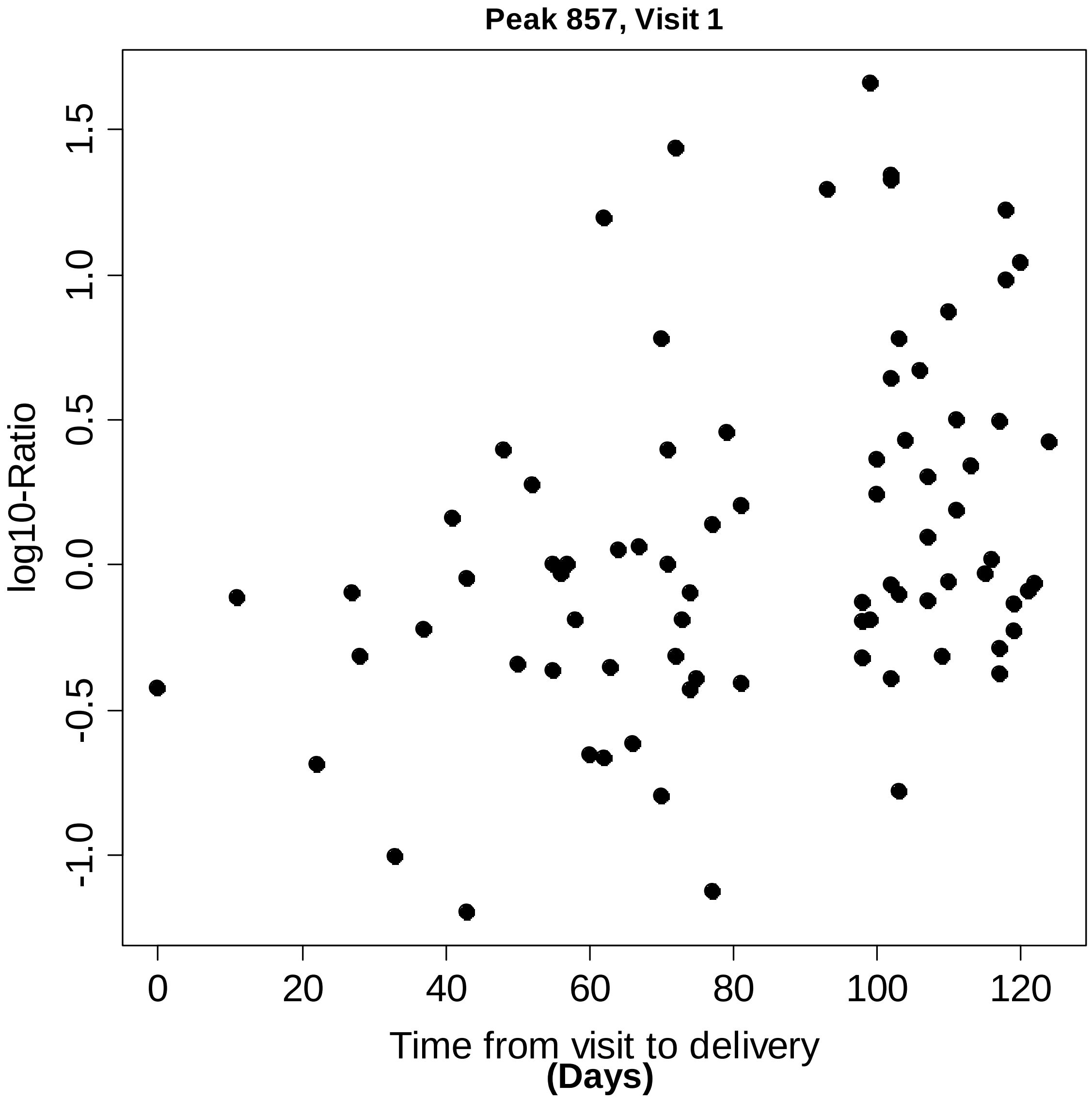


Figure 3.2. A representative plot of one biomarker's normalized abundance (m/z 857) as a function of time to delivery (days). The correlation was statistically significant (R^2=0.11, p=0.003).

PTB, the mean gestational age at delivery being 32.3±0.28 weeks. When biomarker abundance was plotted as a function of time to delivery, a significant correlation was found for all three markers at 28 wks gestation (peak 677: R^2=0.13, p=0.001, peak 857: R^2=0.11, p=0.003, peak 860: R^2=0.12, p=0.002) and for two of the markers at 24 wks gestation (peak 857: R^2=0.13, p=0.001, peak 860: R^2=0.11, p=0.003). See Figure 3.2 for a representative plot. Correlation for the third marker (peak 677) at 24 wks had a p-value of 0.08 (R^2=0.04). In each case abundance of the biomarkers was lower the nearer the delivery. None of the markers demonstrated a correlation between its abundance and gestational age, as would be expected given the narrow timing of specimen collection.

Relationship of Biomarker Abundance as a Function of Etiology of PTB

In the subjects sampled at 24 weeks and 28 weeks, chorioamnioitis was confirmed in only 4 of 80 subjects and 2 of 80 subjects respectively. Levels of all 3 biomarkers were markedly reduced in women with confirmed chorioamnioitis, but the number of subjects having this diagnosis was too small for meaningful statistical comparisons. There was no reduction in the abundance of any of the biomarkers with fetal fibronectin positivity.

Discussion

We have applied a serum proteomics method utilizing cLC-ESI-TOFMS to the analysis of sera collected from pregnant women at 24 and 28 wks gestation. We have identified 3 peptides within the serum of pregnant women at both 24 and 28 weeks gestation that are significantly decreased in women who experienced a later SPTB. The changes predated on average the SPTB by 7.8 wks and 4.7 wks respectively. All three identified peptides came from a single protein and from a highly conserved proline-rich region of that protein that had been processed differently. One of these peptides had an oxidized methionine. The parent compound is termed inter-alpha trypsin inhibitor heavy chain 4 (ITIH4), a glycoprotein that is a kallikrein-sensitive acute phase reactant (Piñeiro

2004). The intact protein is known to be increased in inflammatory states (Piñeiro 2004), but little is known about the function of this protein or its peptide fragments, including possible biological activity. Peptides derived from ITIH4 that differ from the peptides described here have demonstrated quantitative increases in sera of women with early-stage ovarian cancer (Song 2006). Other peptides arising from this same protein, but differing from the peptides described here, appear to be increased in other cancers in a disease specific manner (Song 2006). This might suggest differential peptide production having a disease specific pattern.

Early efforts to carry out serum proteomics resulted in methodological controversies. (Petricoin 2002 Lancet, Baggerly 2004) First, the use of computers to evaluate mass spectral data is challenging. No software application has been accepted as fully reliable. Hence, continual reference to the actual mass spectra is critical. Second, day to day variability can be substantial for both the separation step and subsequent mass spectral analysis. Much of the variability can be eliminated or dramatically decreased by means of an internal control. Such standards are often used to correct for inconsistencies in elaborate, multi-step analyses involving human serum. In this study, such controls were employed by utilization of endogenous reference molecules present in all specimens to compensate for variability in specimen processing, chromatography loading and separation, ionization efficiency and instrumental performance.

Collectively, these data suggest strongly that the three peptides described here are useful biomarkers, identifying approximately two thirds of pregnant women who will delivery prematurely, weeks prior to the SPTB. These significant quantitative differences were observed for both sets of women with later SPTB sampled at different gestational ages and were also significantly different for other charge states of the same peptides as observed by MS. In addition, the data demonstrated a significant relationship between biomarker abundance and nearness to delivery, abundance being lower the closer the SPTB. The sensitivity of the most discriminating peptide or for combinations of the 3 was 65% with a specificity of 82.5% (LR+=3.71, LR-=0.42). (Figure 3.1 C) However, the data suggest that these markers may become better predictors of SPTB as women near the event. It is also highly likely that additional biomarkers will be found that increase the sensitivity and specificity of SPTB prediction. For example, when the current three peptides were coupled to 5 previously tested candidate biomarkers, the sensitivity was 89.8% with a specificity of 81.0% (LR+=4.4, LR-=0.17).

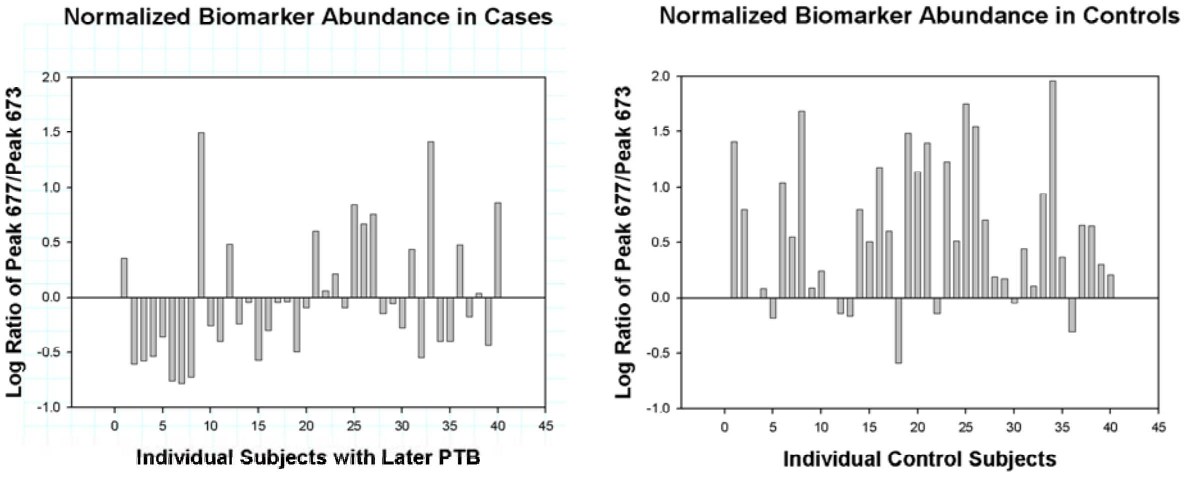


Figure 3.1 C. Normalized biomarker abundance was calculated for all individuals at 28 weeks. The graph shows biomarker (m/z 677) abundance in the sera of women having a later SPTB or term delivery. Sixty five percent (26/40) of the women with later SPTB had a value below the threshold whereas only 17.5 (7/40) percent of women with a term delivery had a value below the same threshold.

The data do not allow for any confident differentiation of the biomarkers according to a potential etiology for the SPTB. However, we hypothesize that women with infectious etiology of their preterm birth will have lower levels of these 3 markers. This will be the focus of future investigations.

The specimens utilized in this analysis were part of a multi-center study representing 10 medical centers located across the US and represent a good mix of subjects. However, it is recognized that these findings will need to be confirmed in a larger number of specimens, preferably in a prospective fashion. In addition, studies of active preterm and term labor are needed to define whether these changes are observed only 4-8 wks prior to a preterm delivery or whether they are still present at the time of active preterm labor and whether the changes in biomarker abundance are limited to preterm births or also accompany term labor also.

Chapter 4 – Serum Biomarkers that Allow Prediction of Preeclampsia in Pregnant Women at 12-14 Weeks Gestation.

Summary

Using a unique serum proteomics approach we have found 29 molecules that are statistically quantitatively different in the serum of pregnant women at the end of their first trimester that who will later develop preeclampsia (PE), a life threatening complication of pregnancy. Five of these are sufficiently different that they each appear to have utility in risk assessment of PE 3-6 months prior to there being any clinical manifestations of PE. Seventeen of these markers appear to have utility in risk assessment in weighted combinations for this same patient population. The polypeptide sequences of two molecular species are provided. In addition, information is provided that uniquely identifies five of these molecular species chromatographically without knowledge of the polypeptide sequence.

Background

Preeclampsia is currently defined by an elevation in blood pressure (>140/90 mm Hg) and protein in the urine (>300 mg/24 hr) occurring in the second half of pregnancy in a women without a history of high blood pressure, kidney disease, diabetes or other significant disease. Preeclampsia can include many other abnormalities and having one or more these suggests severe preeclampsia.

Preeclampsia is estimated to affect 3-5% of all pregnancies in the U.S. and is one of the leading causes of maternal death associated with pregnancy in the U.S. and is estimated to result in ~75,000 maternal deaths each year worldwide.

Currently there are no accepted biochemical markers for either the prediction or diagnosis of PE. However, knowledge that a pregnant woman was at high risk for the later development of PE or was at very low risk for the development of PE would be useful to the clinician in terms of the frequency and intensity of antenatal surveillance and potentially in therapeutic interventions for the mother or fetus.

Methods

Patient Population

This study involved 55 pregnant women having blood withdrawn between 12 and 14 weeks of pregnancy who were followed through the completion of their pregnancy. Twenty seven of these women had uncomplicated pregnancies with no evidence of preeclampsia (PE) including no increase in blood pressure or abnormal levels of protein in their urine. These constitute the control group. Twenty eight of these women developed later PE, each after 24 weeks of pregnancy. These women constitute cases. The sera of these 55 women were studied using our proteomics approach.

Sample Preparation

This method has been described in detail in Chapter 2

cLC-MS Analysis

Capillary liquid chromatography (cLC) to fractionate or separate peptides and proteins was performed using a 1 mm (16.2 μL) microbore guard column (Upchurch Scientific, Oak Harbor, WA) and a 15 cm x 250 um i.d. capillary column. The guard column was dry-packed and the capillary column was slurry packed in-house using POROS R1 reversed-phase media (Applied Biosystems, Framingham MA). Column equilibration and chromatographic separation was performed using an aqueous phase

(98% HPLC grade H_2O, 2% acetonitrile, 0.1% formic acid) and an organic phase (2% H_2O, 98% acetonitrile, 0.1% formic acid). Separation was accomplished beginning with a 3 min equilibration at 95% aqueous solution, followed by a 2.75%/min organic phase increase to 60% organic phase, which was increased at 7%/min to 95% organic phase. The gradient is held at 95% organic phase for 7 min to elute the more hydrophobic components of the sample, and then the gradient is returned to 95% aqueous phase over 5 min and held at this concentration for 2 min to re-equilibrate the column. Separations are performed at a flow rate of 5 μL/min. The cLC uses an LC Packings Ultimate Capillary HPLC pump system, with a FAMOS® autosampler (Dionex Corporation, Sunnyvale, CA) controlled by Analyst QS® software (Applied Biosystems, Foster City, CA).

Electrospray-Ionization, Time-of-Flight Mass Spectrometry

Effluent from the cLC was directed into a QSTAR® Pulsar i quadrupole orthogonal time-of-flight mass spectrometer through an IonSpray source (Applied Biosystems). Mass spectra were collected every second for m/z 500 to 2500 from 5 to 55 min elution. Data collection and preliminary formatting were accomplished using Analyst QS® software with BioAnalyst add-ons (Applied Biosystems). Specimens from cases and controls were analyzed together in randomized order. This is described more fully in Chapter 2, Part 3.

Peak Alignment

Because samples run on different days and columns can vary in elution time, 10 endogenous molecular species of average abundance that elute at approximately 2 minute intervals throughout the chromatogram were determined (Table 2.2). This was detailed in Chapter 2, Section 2.

Evaluation of Candidate Molecular Species

Each species that appeared quantitatively different between groups was further evaluated. To facilitate this, a second peak near the candidate peak (in both the chromatographic and m/z dimension) but quantitatively comparable as determined by visual inspection of a sampling of individual mass spectra in both groups was selected as an internal control. This reference peak was then used to normalize the candidate peak of interest, correcting for variability in specimen processing, specimen loading, ionization efficiency and instrument performance and allowing comparison across runs performed on different days. Thereafter, the molecular species were 'extracted' by the Analyst® software to determine a quantitative peak height of both candidate and reference ions in each specimen for all subjects.

Mass Spectral Data Analysis

The abundance (peak height) of each candidate and reference species was tabulated as was the calculated ratio of each candidate species abundance relative to reference species abundance within each patient. The log of that ratio was also determined because abundance varied substantially. The data were submitted to statistical analysis.

Selection of Reference Peaks for Chromatographic Normalization

Using 2-D and 3-D visualizations of complete serum runs, multiple peaks that eluted at ~2 min intervals were selected as possible reference peaks. The Extract Ion Chromatogram (XIC) function was used to check the chromatographic elution profile of a small m/z range that included the m/z of the possible reference peaks. To be further considered as possible reference peaks, chromatographic elution profiles of the selected peaks had to be relatively narrow (< 2 min), well shaped (close to Gaussian in shape), and having elution profiles that were well resolved and distinguishable from those of other molecular species of similar m/z. Those peaks that best met these criteria were further investigated in serum runs from various individuals prepared and run on different days. Ten peaks that eluted at ~2 min intervals and were ubiquitously present in all serum samples were selected as markers.

Mass of Endogenous Time Reference (daltons)	**Mean Elution Time (min)**
1464.65	14.68
1439.52	17.01
2009.95	18.83
5062.28	21.34
546.31	23.54
545.33	26.12
1046.67	27.60
636.31	32.44
779.52	34.59
1619.07	36.88

Table 4.1. Mass and Elution Time of the Time Alignment Markers The mass and typical elution time of the reference peaks used for time alignment are summarized in the above table.

Statistical Analysis

Data are expressed as means ± 1 standard error. Species that appeared to be quantitatively different were considered. The twenty nine peaks reported here were those that had p-values ≤ 0.05 upon further evaluation. Those species with the lowest p-values are those that showed the best independent discrimination between case and control

$$class = sign\left[\left(\sum_{i\,\in Peaks} \omega_i I_i\right) + c\right]$$

Equation 1. Equation used to determine case/control assignments using multiple peaks. Sign is a function that returns +1 if the equation value is positive and -1 if the equation value is negative. Class is a variable. +1 is returned if the sample is a case, and -1 if the sample is a control. The intensities of each peak are multiplied by a weight and a constant is added. We then look to see how well cases and controls are separated, with the line drawn between them at 0.

samples and are reported in Table 4.4. All of the 29 'significant' peaks were analyzed in combination with each other for optimum discrimination between case and control peaks. Combinations of as few as two peaks and as many as six peaks allowed for discrimination between cases and controls with specificities ≥ 96% and sensitivities ≥ 71%. Case/control determinations were determined using Equation 1 where ω is the relative weight assigned to that peak within the combination, I is the XIC peak intensity for a particular peak within each individual sample, and c is a constant for that combination.

cLC-MS/MS Analysis

Candidate species demonstrating statistically different abundances between cases and controls were further analyzed in an effort to chemically identify the candidate molecules. A frozen specimen of supernatant (from the protein reduction step) containing 0.5 μg of total protein was thawed and loaded onto the column and a MS run was performed to determine the exact elution time of the peptide of interest. A MS/MS run was performed on the same amount of protein with collection of Multi-Channel Acquisition (MCA) fragmentation data at a set collision energy for the two min time span in which the peptide of interest eluted. MS/MS fragmentation data were produced at several discrete collision energies, to produce fragmentation spectra that covered as much of the total peptide sequence as possible. The spectrum were summed and visually compared to exported data lists to insure that the software had assigned charge states correctly. After any correction, the data list was transformed using the formula: +1 mass = m/z value * charge – (charge – 1) so that all species had a +1 mass listed. The corrected list was submitted to Mascot (www.matrixscience.com), a searchable MS database allowing protein/peptide identification. The first peptide identified had a monoisotopic mass of 2027.1 and was successfully identified as a fragment of inter-alpha trypsin inhibitor heavy chain 4, and the second peptide had a monoisotopic mass of 1237.5 and was successfully identified as a fragment of fibrinogen B beta chain that was amidated at the C-terminal end and had a pyroglutamic acid at the N-terminal end.

Results

Biomarker Characteristics

After time alignment, biomarker candidates were identified visually in an initial process where multiple mass spectra were overlaid with cases and controls each assigned a color. Those peaks that appeared to be predominantly one color were studied further. The individual spectra were then submitted to peak height determination by the computer equipped with Analyst® software (Applied Biosystems) which is the operating system for the QqTOF mass spectrometer (Applied Biosystems). The XIC intensity of each of the potential biomarkers was then tabulated. In addition a second peak that occurred in the same time window which was not quantitatively different between cases and controls was also selected. This represented an endogenous control for each peak under investigation, allowing for reduction of non-biologic variability. This was accomplished by dividing the quantity of the candidate peak by the quantity of the endogenous control. The magnitude of this ratio for each specimen was recorded and statistical differences were sought using a Student's t-test comparing cases and controls.

	Peak (m/z)	**Mean Mass (daltons)**	**Mean Elution Time (min)**
1.	718.8	4305.943 ± 0.020	20.40 ± 0.83
2.	719.2	4313.199 ± 0.118	20.24 ± 0.77
3.	734.8	1647.506 ± 0.022	19.40 ± 1.42
4.	649.3	648.322 ± 0.037	24.27 ± 0.67
5.	507.3	506.306 ± 0.011	17.64 ± 0.67

Table 4.2. Mass and elution times of 5 prospective biomarkers for PE

While currently 29 molecular species have been found to be statistically different, five of these were sufficiently different ($p \leq 0.0001$) to suggest that they might allow for excellent separation of the two groups without combination of any other peak. The individual masses and elution times for the five PE biomarkers are summarized in Table 4.2.

The elution time (retention time) can also be expressed as a function of the internal time controls. This is determined by the relative position of the peak of interest between the time marker that runs precedes the biomarker and the time marker that follows the peak of interest. This can be calculated as:

$\mathbf{R_f}$ = (elution time of biomarker – elution time of preceding time marker) / (elution time of following time marker – elution time of preceding time marker)

The R_f values are more reliable than the actual elution times. Elution times may vary with new columns or with the altered performance of an existing column due to fouling, but the R_f is not altered by these changes. The R_f values of the five biomarkers are provided in Table 4.3.

	Peak (m/z)	**N**	**R_f Value Relative To Boundary Time Markers**
1.	718.8	12	0.635 ± 0.85
2.	719.2	12	0.737 ± 0.072
3.	734.8	9	0.294 ± 0.024
4.	649.3	10	0.343 ± 0.120
5.	507.3	11	0.359 ± 0.039

Table 4.3. The R_f Values for the PE Biomarkers Using the Internal Time Alignment Peaks.

Reduction of Variability by Reference to an Endogenous Co-eluting Control

One of the features of the current serum proteomic approach is the use of an endogenous molecule that is found in all species and is not different between cases and controls. Normalization of biomarker abundance to this internal control reduces non-biological variation and improves the ability to utilize biomarkers in risk prediction. Normalization involves mathematically dividing the abundance of the peak of interest by the reference peak. The abundances are machine derived values. The abundance of a given molecule represents the number of ions of a particular mass measured by the mass spectrometer in a given mass spectrum or the sum of the number ions of a specific mass observed in several mass spectra representing the full elution interval. Molecules typically require 1.0 -1.5 min to move off the chromatographic column whereas mass spectra are acquired every 1 second during that elution interval.

The first two peaks with m/z 718.8 and 719.2 are in fact both significantly different between cases and controls but the first is more abundant in cases and the second is more abundant in controls. These two peaks are simply referenced to each other, i.e. the abundance of the m/z 718.8 is divided by the abundance of the m/z 719.2. For the other three peaks internal references were used. For the biomarker peak m/z 734.8, a coeluting reference peak at m/z 725.4 was chosen. For the biomarker m/z 649.3, a coeluting reference peak at m/z 512.3 was chosen. For the biomarker m/z 507.3, a coeluting reference at m/z 734.5 was chosen.

Using these ratios the mean value for the log ratios are as follows (Table 4.4):

Ratio	Mean Control	Mean PE	P value
1. log 718.8/719.2	-0.440 ± 0.205	-0.0788 ± 0.255	2 x 10^{-7}
2. log 734.8/725.4	-0.278 ± 0.225	-0.0215 ± 0.123	0.000003
3. log 649.3/512.3	-0.098 ± 0.386	+0.315 ± 0.323	0.00003
4. log 507.3/734.5	+0.400 ± 0.524	-0.0944 ± 0.3962	0.0001

Table 4.4. Biomarker Abundance (after Normalization) in Cases and Controls

This method for normalization to internal controls was performed for all 29 potential biomarkers. These results are summarized in Table 4.5. The top 17 out of 29 peaks provided more powerful predictive capabilities when combined with other peaks.

m/z	Ref Peak	charge	Mono-isotopic mass	p value	color
734.8	742.8	2	1468.6	3.0E-06	🟨
507.2	734.5	1	507.2	1.5E-04	🟥
639.3	582.3	1	639.3	1.0E-05	⬜
1026.4	518.3	2	2051.8	1.0E-04	🟦
942.5	559.2	1	942.5	2.4E-02	🟩
649.3	512.3	2	1297.6	3.0E-05	🟦
634.3 L	582.3	1	634.3	7.0E-04	🟪
1238.5	623.4	1	1238.5	3.2E-02	🟩
619.8	623.4	2	1238.6	4.8E-02	🟧
719.2	725.4	6	4310.2	1.0E-03	🟩
718.9	725.4	6	4308.4	5.0E-02	🟧
553.3	518.3	1	553.3	1.0E-04	⬛
739.4	518.3	2	1477.8	1.0E-04	🟦
634.3 R	582.3	1	634.3	2.0E-04	🟦
1074.4	518.3	2	2147.8	4.0E-04	⬛
509.3	649.4	2	1017.6	1.0E-03	🟫
505.3	649.4	2	1009.6	7.0E-03	🟪
718.9	719.2	6	4308.4	2.0E-07	🟪
757.2	734.8	3	2269.6	4.0E-03	(white)
571.3	565.3	1	571.3	8.0E-03	none
585.3	579.3	1	585.3	9.0E-03	none
541.3	512.3	2	1081.6	1.1E-02	none
601.3	512.3	2	1201.6	1.2E-02	none
676.7	685.2	3	2028.1	1.9E-02	none
785.3	512.3	1	785.3	1.9E-02	none
756.4	758.5	1	756.4	2.0E-02	none
805.3	559.2	1	805.3	2.8E-02	none
593.3	572.3	1	593.3	3.9E-02	none
783.3	559.2	1	783.3	4.6E-02	none
568.8	565.3	4	2272.2	5.0E-02	none

Table 4.5. Prospective Biomarkers Analyzed. A list of the 29 potential biomarker peaks that were determined to be statistically significant (p-value ≤ 0.05). The top 18 of these were used in combination with each other for the weighted discrimination analysis.

Use of the Biomarkers to Predict Women at Risk of a Later Preeclampsia

One common measure of the predictive power of a biomarker is its sensitivity and specificity. Sensitivity is a statistical term defined as the true positive rate or specifically in this case the percentage of pregnant women who later develop PE that are correctly identified by the biomarker. The specificity is defined as the true negative rate or in this case the percentage of pregnant women with uncomplicated pregnancies correctly identified. To use a biomarker for prediction in this manner a numeric threshold must be established. To establish that numeric value, typically the range of values for the biomarker are considered from lowest to highest and at each point the percent of subjects correctly identified as positive and at that same point the percent of controls incorrectly identified as positive. This is termed a receiver operator curve (ROC). Consistent with general practice the false positive rate is limited to 20%. This is commonly considered the maximum tolerated for a clinical test. The false positive rate (the percentage of women with uncomplicated pregnancies, the control group, identified by the biomarker as at risk for later PE) is calculated from the true negative rate being subtracted from 100%. Whatever the threshold is at a false positive rate of 20% or less (which is equivalent to a specificity of 80% or higher) determines the threshold used to determine whether someone is at risk or is not at risk. A threshold for each of the four ratios was determined that allowed for the identification of subjects at risk of later PE. The threshold for each was calculated such that there would be a specificity (a true negative rate) of 80% or more. As stated, this is the same as a false positive rate no more than 20%. Using these mathematically determined thresholds the

Ratio	**Threshold**	**Sensitivity**	**Specificity**
1. log 718.8/719.2	≥ -0.301	82%	85%
2. log 734.8/725.4	≥ -0.11	71%	85%
3. log 649.3/512.3	≥ 0.253	67%	80%
4. log 507.3/734.5	≤ -0.125	48%	85%

Table 4.6 Sensitivity and Specificity of Each Biomarker (after Normalization)

four ratios independently provided the following sensitivity (true positive) and specificity (true negative) rates for single peptide biomarkers as summarized in Table 4.6.

If one combines the log 718.8/719.2 ratio (at a threshold of > -0.301) with the ratio of log 649.3/512.3 now with a more stringent threshold of > 0.301, the sensitivity improves to 89.3% with a specificity of 85%. Combination of the log 718.8/719.2 ratio with log 734.8/725.5 with a threshold > -0.10 provides a sensitivity of 100% with a specificity of 74%. Most of the potential biomarker peaks we determined to be statistically significant do not result in high enough sensitivities or specificities when considered as the sole discriminating factor between case and control sample sets. The sensitivities and specificities were increased when multiple peaks were used in weighted combinations. Various combinations of the peaks were analyzed. 17 of the peaks were used in combination with as few as one and as many as five other peaks to provide a total of 15 combinations that discriminated between the case and control groups with specificities ≥ 96% and sensitivities ≥ 71%. (Table 4.7)

				Biomarkers				Const.	Specificity	Sensitivity
		1	2	3	4	5	6		Out of 23	Out of 24
1	Mass	507.2	734.8	942.5	634.3 L	639.3			100%	83%
	Window	2	3	1	7	7			23/23	20/24
	ω	1	-16	15	4	1			# wrong	4
2	Mass	507.2	619.8	718.9	942.5	649.3	634.3L		96%	88%
	Window	1	3	3	1	4	7		22/23	21/24
	ω	27	1	-32	16	-32	31	4	# wrong	4
3	Mass	507.2	734.8	739.4	639.3				100%	83%
	Window	2	3	6	7				23/23	20/24
	ω	4	-8	1	6				# wrong	4
4	m/z	507.2	619.8	734.8	649.3	505.3	1074.4		96%	88%
	Window	2	3	3	4	5	6		22/23	21/24
	ω	1	1	-2	-2	1	1		# wrong	4
5	m/z	507.2	619.8	734.8	942.5	649.3	639.3		96%	88%
	Window	2	3	3	1	4	7		22/23	21/24

	ω	2	1	-4	1	-4	2		# wrong	4
6	m/z	507.2	1238.5	734.8	649.3	509.3	553.3		100%	83%
	Window	2	3	3	4	5	6		23/23	20/24
	ω	3	2	-4	-4	1	2		# wrong	4
7	m/z	718.9	719.2	734.8	739.4	634.3 L			100%	83%
	Window	3	3	3	6	7			23/23	20/24
	ω	-16	8	-16	3	1			# wrong	4
8	m/z	757.2	719.2	942.5	649.3	634.3 L			100%	83%
	Window	2	3	1	4	7			23/23	20/24
	ω	-1	1	17	-1	4			# wrong	4
9	m/z	507.2	719.2	734.8	649.3	509.3	639.3		96%	88%
	Window	2	3	3	4	5	7		22/23	21/24
	ω	4	2	-8	-8	1	4		# wrong	4
10	m/z	507.2	1238.5	734.8	942.5	505.3	634.3 L		100%	83%
	Window	2	3	3	1	5	7		23/23	20/24
	ω	2	2	-4	2	1	1		# wrong	4
11	m/z	507.2	718.9	942.5	649.3	553.3	1026.4		96%	88%
	Window	2	3	1	4	6	6		22/23	21/24
	ω	1	-1	2	-2	5	9		# wrong	4
12	m/z	734.8	1074.4						100%	71%
	Window	3	6						23/23	17/24
	ω	-1	8						# wrong	7
13	m/z	734.8	1026.4	639.3	12/12.1				96%	100%
	Window	3	6	7	3				22/23	24/24
	ω	-5	33	2	-2				# wrong	1
14	m/z	734.8	1026.4	639.3					87%	100%
	Window	3	6	7					20/23	24/24
	ω	-3	17	1					# wrong	3
15	m/z	734.8	942.5	1026.4	634.3R				96%	96%
	Window	3	1	6	7				22/23	23/24
	ω	-4	2	18	3				# wrong	2
16	m/z	1238.5	734.8	1026.4	639.3				96%	96%
	Window	3	3	6	7				22/23	23/24
	ω	1	-16	64	3				# wrong	2
17	m/z	734.8	942.5	1026.4	634.3R				96%	96%
	Window	3	1	6	7				22/23	23/24
	ω	-2	1	9	2				# wrong	2
18	m/z	734.8	1026.4	634.3R					91%	96%
	Window	3	6	7					21/23	23/24
	ω	-1	6	1					# wrong	3

Table 4.7. Weighed Discrimination Analysis. Combinations of as few as two peaks and as many as six peaks allowed for discrimination between cases and controls with specificities ≥ 87% and sensitivities ≥ 71%. Case/control determinations were determined using Equation 1 where ω is the relative weight assigned to that peak within the combination, I is the XIC peak intensity for a particular peak within each individual sample, and c is a constant for that combination. If class > 0 for a sample, that specimen is assigned to the case set, whereas if class ≤ 0 for a sample, that specimen is assigned to the control set.

Discussion

We have applied the same proteomics approach as was described in Chapter 3 to the analysis of sera collected from pregnant women at 12-14 weeks gestation. The cases (n=23) went on to develop preeclampsia (typically occurring after 24 weeks gestation). The controls (n=24) had normal, healthy pregnancies. Though these sera were collected very early in the pregnancy, we have found 29 molecular species present at statistically different levels between the case and control sample sets. The ratio between two biomarker peptides very similar in both elution time and m/z successfully predict ~82% of women who will later go on to develop PE more than 3 months prior to the onset of symptoms (Figure 4.1, Figure 4.2). Using a several potential biomarkers in a combined

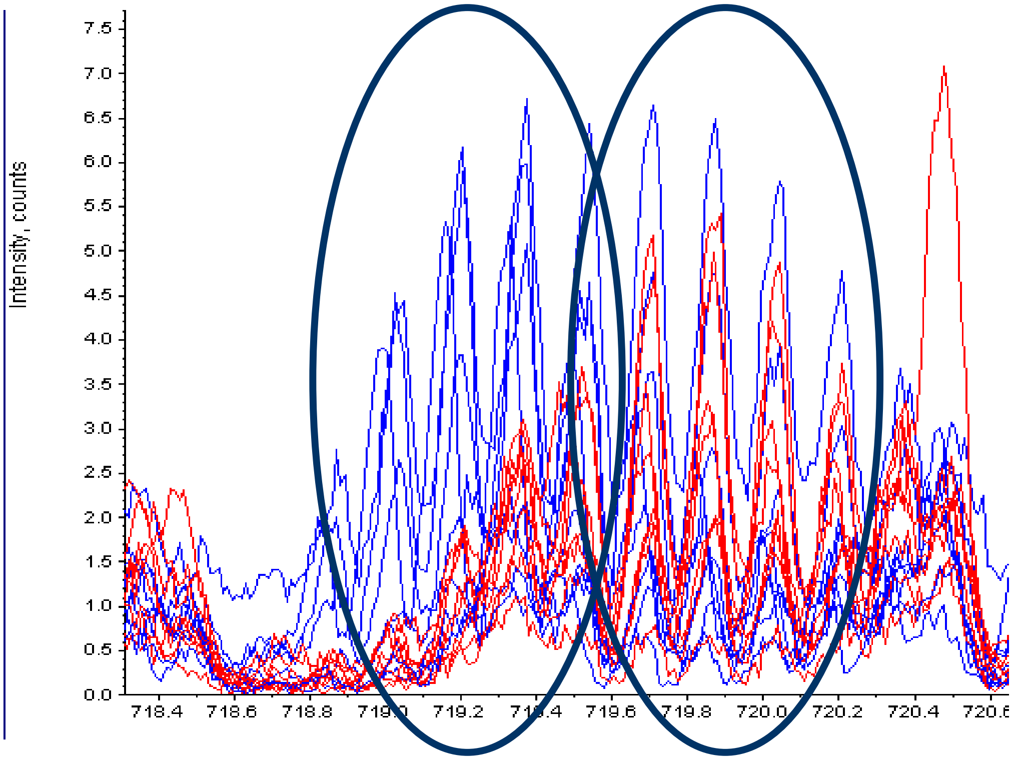


Figure 4.1. Peptide Biomarkers that predict ~82% of women who will go on to develop preeclampsia. A sample overlay of the spectra of 8 women who went on to develop PE (blue) and 8 women who had normal, healthy pregnancies.

classification approach (Table 4.7) increases the sensitivity up to 100% and the specificity up to 96% (Figure 4.3).

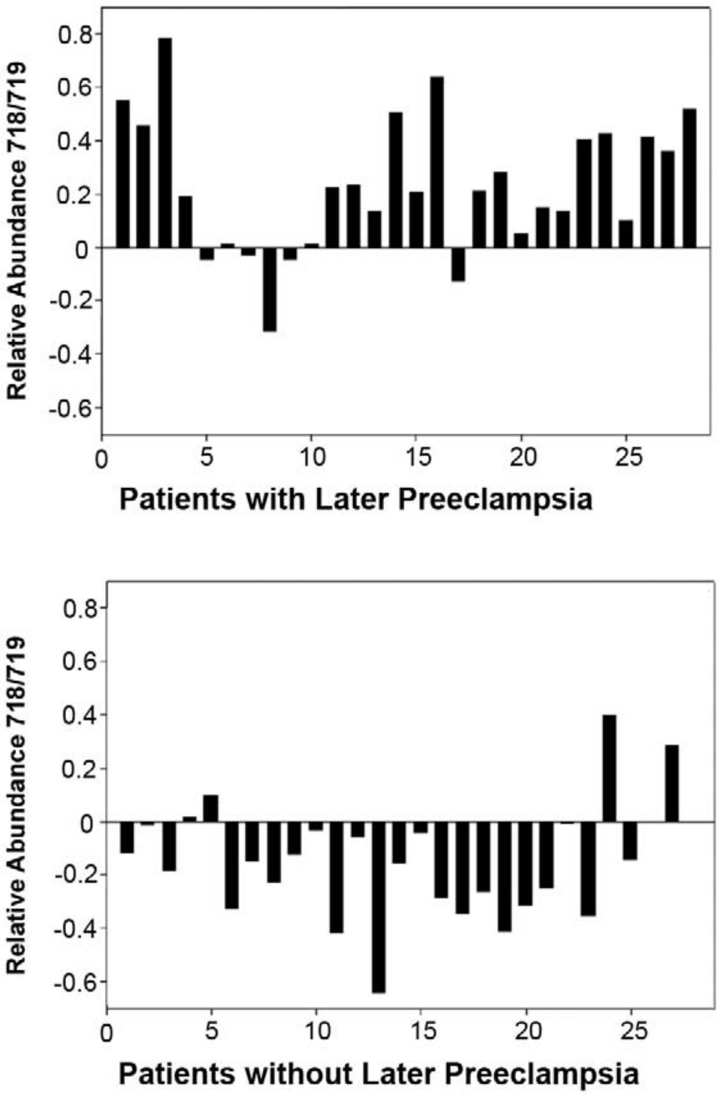


Figure 4.2. Classification of case/control patients using ratio of the relative abundance of peak 718.8/719.2. These two peaks alone allowed for differentiation of the two groups with a sensitivity of 82% and a specificity of 85%. This equates to a misclassification of only 5 out of 28 case samples and 4 out of 27 control samples.

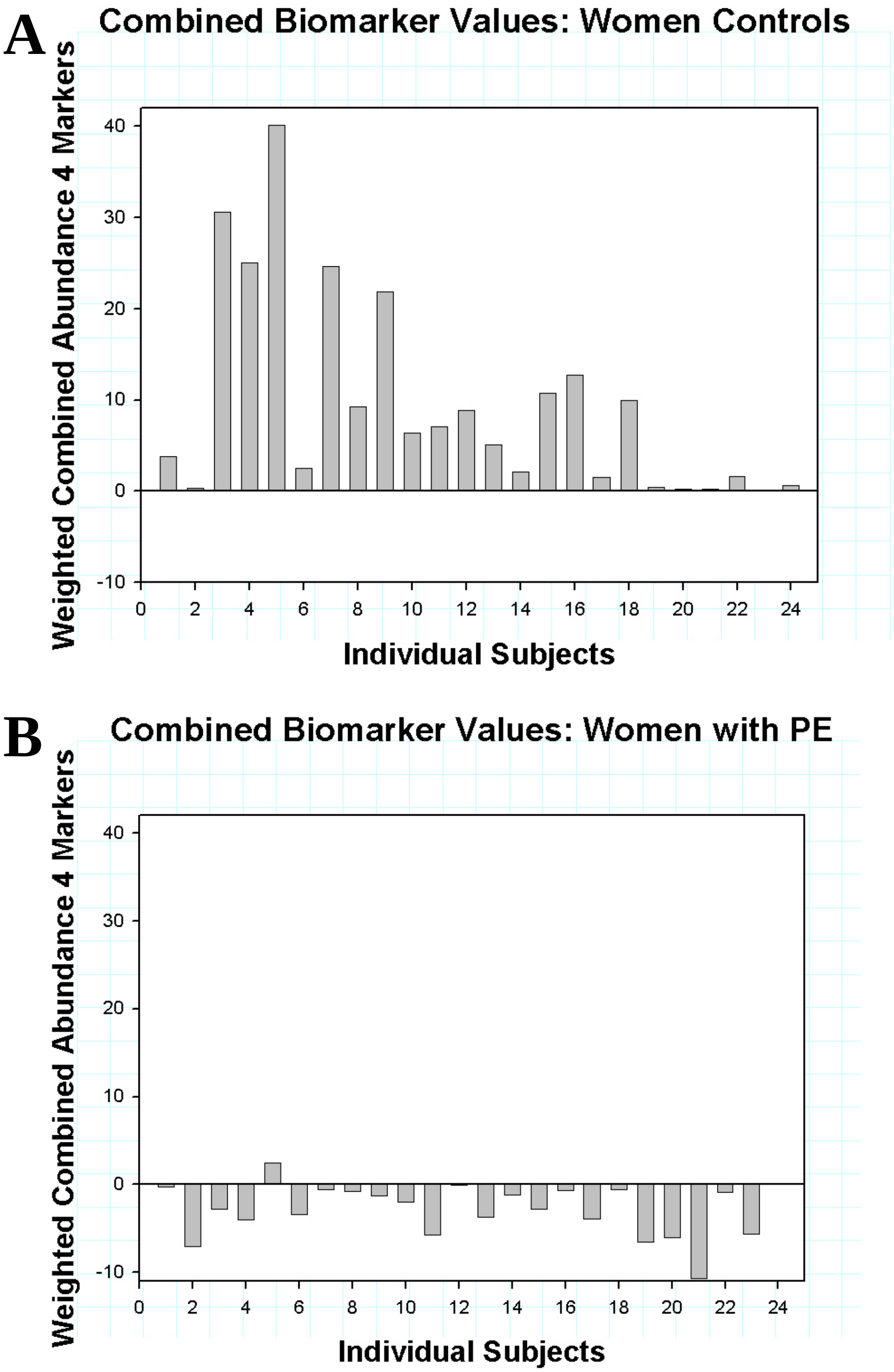


Figure 4.3. Classification of case/control patients using a combination of different peptide ratios. Using a combination of different peaks allowed for differentiation of the samples into case/control groups with a maximum sensitivity up to 100% and a maximum specificity up to 96%. This equates to a misclassification of 1 out of 23 case samples and 0 out of 24 control samples. The markers used were 639.3, 718/719, 734.8, and 1026.4.

Two of the peptides identified as potential biomarkers for preeclampsia have successfully been sequenced. The peak with m/z 676.7 was observed in its +3 charge state and was identified as a fragment of the inter-alpha trypsin inhibitor heavy chain 4 with a polypeptide sequence of (qlglpgppdvpdhaayhpf). ITIH4 is an acute phase reactant that also inhibits actin polymerization and phagocytosis of polymorphonuclear cells. This class of peptides has been observed to increase in patients with inflammatory disorders and may prove to be a mediator of infection or inflammatory response. (Pineiro 2004) The peak with m/z 691.8 was observed in its +2 state and was identified as fibrinogen B beta chain that was amidated at the C-terminal end and had a pyroglutamic acid at the N-terminal end. The polypeptide sequence of this species was (pyro-egvndneegff-NH_2). Fibrinogen is involved in blood clotting, fibrinolysis, cellular and matrix responses, wound healing, the inflammatory response, and neoplasia. (Mosesson 2005) The reason for an increased amount of fibrinogen B beta chain we observed in serum samples of women who developed preeclampsia is unknown, but we can surmise that the peptide is not merely a cleavage fragment. The N-terminal end of this fragment had a pyroglutamate and the C-terminal end was amidated. These modifications are typically mediated by enzymes and suggest that this fragment was made by some cell for a specific purpose.

Collectively, these data suggest strongly that the peptides described here have great potential as useful biomarkers, identifying close to 100% of pregnant women who will later develop preeclampsia. All of the peptides reported here had p-values ≤ 0.05, however, the data suggest that these markers may become better predictors of PE in combination with other markers rather than using individual markers for risk assessment.

It is highly likely that additional biomarkers will be found that increase the sensitivity and specificity of PE prediction. Further sequence identification of these potential biomarkers may give insight to the pathology of this disease. This study was performed using a relatively small sample set (n=55) and we recognize that these findings will need to be confirmed in a larger number of samples. Additional samples can be studied to see if the levels of these peptides change as delivery approaches. Women determined at high risk for development of PE can be closely monitored and additional efforts can be made to prevent onset of this disease.

Chapter 5 – Concluding Remarks

Summary of Current Research Accomplishments

Sample Preparation

We developed a protein precipitation method that effectively removed the large, highly-abundant proteins from serum. We showed that as the organic solvent used in the precipitation protocol denatured the larger molecules, smaller molecules bound to larger 'carrier proteins' were released into the solution. We demonstrated both the reproducibility of this method and the increased detection of smaller proteins and peptides that were present in low amounts in serum. This method facilitated the search for novel biomarkers for disease within the low-abundance, low-molecular weight serum proteome.

Chromatographic Time Normalization

Due to the sensitivity of MS instrumentation, there is an inherent chromatographic variability in the data from day-to-day and even run-to-run. To correct for this variation, we developed a method to use 10 native serum peaks that were found in all samples as internal controls. The 10 peaks selected eluted at approximate 2 minute intervals, and were used as chromatographic reference points to normalize sample data run on different days. Two-minute time windows centered on the elution time of a given marker were overlayed for up to 16 different samples. Since all samples run were that of human serum processed in the same manner, one would expect that the majority of the peptides and proteins present in various serum samples would remain consistent. Without a reliable method for chromatographic time alignment, samples did not appear to be consistent. Using this method of internal standard alignment to overlay the samples, the consistency from sample to sample was greatly increased. This allowed for visual inspection of these overlays for variations between the case and control sample sets. Any peaks appearing to be quantitatively different can then be further analyzed to determine their potential as biomarkers for the disease being investigated.

Sequencing of Intermediate to Large Peptides

The potential for discovery of biomarkers within the LMW fraction of the serum proteome is great. Even after removal of the majority of protein from serum using organic solvent precipitation, several thousand peaks were observable from a cLC-MS run of just 0.5 μg of sample. After time alignment and analysis, peaks of interest were selected and subjected to further analysis to determine quantitative differences between case and control sets. Once these differences were determined, a protein or peptide was ideally sequenced for optimum use as a potential biomarker. While it is possible to adequately describe a unique peak using only m/z and relative elution time, it is impossible to make any useful observations as to the possible role of a biomarker in the pathology or progression of a disease without knowing the amino acid sequence of that peptide. The method described here facilitated the identification of intermediate to large serum peptides in their native, undigested form without the need for further sample purification or processing. Even large, multiply charged peptides that were present in relatively low quantities could potentially be sequenced using this method. As a demonstration of the effectiveness of this method, a peptide with a m/z of 714.2 and a charge state of +6 and a peptide with a m/z of 844.6 and a charge stated of +6 were sequenced from a sample of serum. This method allowed for daughter ion fragment coverage of 33 of 39 of the amino acids for the peptide of m/z 714.2 and 42 out of 47 amino acids for the peptide with m/z of 844.6.

Biomarker Search for Preterm Birth

Serum samples of pregnant women were drawn at 22-24 weeks (visit 1) and 26-29 weeks (visit 3) gestational age. Some of these women went on to deliver their babies pre-term (<37 weeks) and were considered the case subjects. The other women all went on to deliver healthy, term babies and were considered to be the control set. The sample set size analyzed was n=40 for each of the four sample categories (V1 case, V1 control, V3 case, V3 control) The methods outlined above were used to run the serum samples and search the spectra for potential biomarkers. Potential biomarkers that were determined to effectively discriminate between case and control samples (p-value < 0.05) were then subjected to fragmentation in an effort to identify their protein sequence. For this sample set, 3 peptide biomarkers were fully identified. A combination of these three biomarkers predicted ~67% of women that went on to have a PTBup to 8 weeks prior to the actual onset of preterm labor. A combination of these 3 biomarkers with 5 previously studied candidate biomarkers provided a specificity of 89.8% and sensitivity of 81.0 %. Five additional candidate serum markers were found at significantly altered levels, but have not been successfully sequenced. Prevention is key to avoid PTB, and use of these novel biomarkers for risk-assessment months in advance could have a positive effect on the delay of preterm labor.

Biomarker Search for Preeclampsia

Serum samples of pregnant women were drawn at 12-14 weeks gestational age and tracked through the end of their pregnancy. Some of these women (N=27) went on to have normal, healthy pregnancies, and the others (N=28) went on to develop preeclampsia later on in their pregnancies. These sera were processed, run, and analyzed as previously discussed. The spectra were then normalized, overlaid, and inspected for the presence of potential serum biomarkers. A total of 29 unique molecular species were found that were present at statistically significant levels between the case and control sample sets. A ratio of two of the more promising peaks discriminated between the groups with a p-value of 2.0×10^{-7} and successfully predicted ~82% of women having PE more than 3 months prior to clinical symptoms. The discriminatory power of the 29 candidate biomarkers were tested in weighted combinations. The best of these weighted combinations provided 96% sensitivity and 100% specificity more than 3 months prior to clinical disease. Two of the 29 biomarkers have been fully identified using MS/MS.

Limitations of Current Research

Serum Protein Subset

In order to assure the release of LMW peptides from larger carrier proteins in serum, and to assure reproducible protein depletion, two volumes of organic solvent were added to one volume of raw serum. This caused almost all proteins > 5kDa to precipitate out of the solution. This small subset of serum proteins has proven to show promise in the search for novel serum biomarkers for disease, but we recognize that there are most likely several potential serum markers > 5kDa in other molecular weight ranges that we are not able to observe. Future efforts may attempt to characterize a different molecular weight subset of the serum proteome.

Sample Set Limitations

Samples for the PTB study were supplied by the MFMU network of the NICHHD, collected as part a multi-center study. Having many research centers all over the US contribute samples assured a good sampling of a wide demographic of women. These samples were collected between 1992 and 1994. Long-term effects on the storage of serum samples at -80ºC have not been thoroughly studied. All of these samples have been stored under identical conditions so any variability introduced by storage is less likely to affect studies comparing case and control from this set. Comparisons of these samples to those that may have been collected more recently may prove to be less reliable. Samples for the PE study were collected at local hospitals and so do not represent as wide of a demographic as the PTB samples. Only enough serum for one preparation (200 µl) was provided for the PE study. These samples had a much shorter storage time than those for PTB, but small sample volumes limited how many runs can be made to try and sequence peaks of interest.

Analysis of Data

A single MS run used 0.5 µg of total protein and eluted it over a 55 minute gradient. The raw text data file for one sample can reach up to a gigabyte in size. The large size of these files and the complexity of the data for samples of this type makes data analysis very difficult. There are no software packages commercially available that can adequately perform the alignment, normalization, deconvolution of charge states, and comparison of the dozens (and potentially hundreds) of serum MS runs necessary in the search for potential biomarkers. The current method for analysis (described in Chapter 2) involved overlay of 8 case and 8 control spectra at a time and visually inspecting for differences in case and control samples that are further inspected as biomarker candidates. This process is laborious and time-consuming, but has proven the most reliable method for finding prospective biomarkers.

Future Research Objectives

Identification of Additional Biomarkers

For the PTB data set, there is still ~50% of the MS data yet to be analyzed. There may be many more potential biomarkers identified from this yet unanalyzed portion of the data. In addition, there are over 200 more samples from this same specimen set that have been collected, but not analyzed at all. These samples could be used to search for biomarkers, or as part of a study to validate the potential biomarkers singled out by the initial set of 160 samples. The peptide sequence and protein identification has been determined for only 3 biomarkers for the PTB set and 2 for the PE set. Many more peptides have been selected and statistical analysis has shown that they can discriminate between case and control samples with p-values of less than 0.05. To maximize the utility of these molecular species as biomarkers for disease, the peptide sequence and parent protein identity must be determined. Additionally, knowing which proteins these biomarkers originate from may also give some clues to the pathology or progression of the disease being studied.

Analysis of Additional Ob/Gyn-Related Diseases

In addition to analyzing more samples for PE and PTB to either validate markers already identified or to search for additional potential biomarkers, we may seek to analyze other obstetrically or gynecologically related diseases. Some examples of potential diseases are recurrent miscarriage which affects ~5% of pregnancies, or endometriosis which affects ~10% of women in North America and is a leading cause of infertility and hysterectomy in the US. The causes for these conditions are not well understood, and currently there are not biomarkers to test for these conditions. To complicate matters somewhat, surgery is currently required for a definitive diagnosis of endometriosis. Serum biomarkers for the diagnosis of endometriosis would benefit many women if they would not need surgery to reach that diagnosis.

References

Abersold, R. H., Leavitt J., Saavedra R. A., Hood L.E., & Kent S.B. (1987). Internal amino acid sequence analysis of proteins separated by one- or two- dimensional gel electrophoresis after in situ protease digestion on nitrocellulose. *Proc Natl Acad Sci* **84** (20):6970-6974.

Abersold R., & Mann M. (2003). Mass spectrometry-based proteomics. *Nature* **422** (6928):198-207.

Adkins J. N., Varnum S. M., Auberry K. J., Moore R. J., Angell N. H., Smith R. D., Springer D. L., & Pounds J. G. (2002). Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry. *Mol Cell Proteomics* **1** (12):947-55.

Ahmed N., Barker G., Oliva K., Garfin D., Talmadge K., Georgiou H., Quinn M., & Rice G. (2003). An approach to remove albumin for the proteomic analysis of low abundance biomarkers in human serum. *Proteomics* **3** (10), 1980-1987.

Alpert A., & Shukla A. Precipitation of Large, High-Abundance Proteins from Serum with Organic Solvents. *ABRF 2003 Translating Biology Using Proteomics and Functional Genomics* **Poster# P111-W** (2003).

Anderson N. L., Anderson N. G. (1998). Proteome and proteomics: New technologies, new concepts, and new words. *Electrophoresis* **19** (11):1853-1861.

Anderson N. L., & Anderson N. G. (2002) The human plasma proteome: history, character, and diagnostic properties. *Mol Cell Proteomics* **1** (11):845-867.

Araujo R. P., Petricoin E. F., & Liotta L. A. (2008). Cricital dependence of blood-borne biomarker concentratios on the half-lives of their carrier proteins. *J Theoretical Biol* **253** (3):616-622.

Baar S. J. (1956). Studies on urinary peptides isolated from patients suffering from burns. *J Clin Pathol* **9** (2):144-147.

Baggerly K. A., Morris J. S., Coombes, K. R. (2004). Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments. *Bioinformatics* **20** (5):777-785.

Bailes J., Soloviev M. Affinity peptidomics approach to protein detection, quantification, and protein affinity assays: application to forensics and biometrics. In: Soloviev M., Shaw C., Andrén P., editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Banez, L. L., Prasanna P., Sun L., Ali A., Zou Z., Adam B. L., McLeod D. G., Moul J. W., & Srivastava S. (2003). Diagnostic potential of serum proteomic patterns in prostate cancer. *J Urol* **170** (2):442-6.

Bliss M. The Discovery of Insulin. Chicago, University of Chicago Press, 1982, p155.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther* **69** (3):89-95.

Bogdanov B., Smith R. D. (2005). Proteomics by FTICR mass spectrometry: top down and bottom up. *Mass Spectrom Rev* **24** (2):168-200.

Buhimschi C. S., Bhandari V., Hamar B. D., Bahtiyar M. O., Zhao G., Sfakianaki A. K., Pettker C. M., Magloire L., Funai E., Norwitz E. R., Paidas M., Copel J. A., Weiner C. P., Lockwood C. J., & Buhimschi I. A. (2007). Proteomic profiling of the amniotic fluid to detect inflammation, infection, and neonatal sepsis. *Plos Med* **4** (1):84-94.

Buhimschi I. A., Christner R., & Buhimschi C. S. (2005). Proteomic biomarker analysis of amniotic fluid for identification of intro-amniotic inflammation. *Brit J Obstet Gynaecol* **112** (2):173-81.

Catalona W. J., Smith D. S., Ratliff T. L., Dodds K. M., Coplen D. E., Yuan J. J., Petros J. A., & Andriole G. L. (1991). Measurement of prostate-specific antigen in serum as a screening test for prostate cancer. *N Engl J Med* **324** (17):1156-61.

Cho W. C., Yip T. T., Yip C., Yip V., Thulasiraman V., Ngan R. K., et al. (2004). Identification of serum amyloid A as a potentially useful biomarker to monitor relapse of nasopharyngeal cancer by serum proteomic profiling. *Clin Cancer Res* **10** (1):43-52.

Cutillas PR. Quantification of polypeptides by mass spectrometry. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

de Jong D., Jansen R. W., Kremer B. P., & Verbeek M. M. (2006). Total serum cholesterol and recovery from disability among hospitalized older adults. *Gerontol A Biol Sci Med Sci* **61** (7):755-758.

Diamandis E. (2003). Point-Proteomic patterns in biological fluids: Do they represent the future of cancer diagnostics? *Clin Chem* **49** (8):1272-1275.

Diamandis, E. P. (2004) Mass spectrometry as a diagnostic and a cancer biomarker discovery tool. *Mol Cell Proteomics* **3** (4)*,* 367-378.

Dorsey E. R., Holloway R. G., & Ravina B. M. (2006). Biomarkers in Parkinson's disease. *Expert Rev Neurother* **6** (6):823-831.

Drake R. R., Cazares L., & Semmes O. J. (2007). Mining the low molecular weight proteome of blood. *Proteomics Clin Appl* **1** (8):758-768.

Edman P. (1970). Sequence determination. *Mol Biol Biochem Biophys* **8**:211-255.

Finch P, Soloviev M. Selective delpletion and enrichment methods for the analysis of protein and peptide pools. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Georgiou H. M., Rice G. E. & Baker M. S. (2001). Proteomic analysis of human plasma: failure of centrifugal ultrafiltration to remove albumin and other high molecular weight proteins. *Proteomics* **1 (**12),1503-6.

Goldenberg R. L., Rouse D. J. (1998). Medical progress: prevention of premature birth. N Engl J Med **339** (5):313-320.

Goldenberg R. L., Iams J. D., Mercer B. M., Meis P. J., Moawad A., Das A., Miodovnik M., Vandorsten P. J., Caritis S. N., Thurnau G., & Dombrowski M. P. (2001). Maternal-Fetal Medicine Units Network. The Preterm Prediction Study: toward a multiple-marker test for spontaneous preterm birth. *Am J Obstet Gynecol* **185** (3):643-51.

Goncalves A., Esterni B., Bertucci F., Sauvan R., Chabannon C., Cubizolles M., et al. (2006). Postoperative serum proteomic profiles may predict metastatic relapse in high-risk primary breast cancer patients receiving adjuvant chemotherapy. *Oncogene* **25** (7):981-989.

Good D. M., Wirtala M., McAlister G. C., & Coon J. J. (2007). Performance characteristics of electron transfer dissociation mass spectrometry *Mol Cell Proteomics* **6** (11):1942-1951.

Govorukhina N. I., Keizer-Gunnink A., van der Zee A. G., de Jong S., de Bruijn H. W., & Bischoff R. (2003). Sample preparation of human serum for the analysis of tumor markers. Comparison of different approaches for albumin and gamma-globulin depletion. *J Chromatogr A* **1009** (1-2), 171-178.

Gravett M. G., Novy M. J., Rosenfeld R. G., Reddy A. P., Jacob T., Turner M., McCormack A., Lapidus J. A., Hitti J., Eschenbach D. A., Roberts C. T. Jr., & Nagalla S. R. (2004). Diagnosis of intra-amniotic infection by proteomic profiling and identification of novel biomarkers. *JAMA* **292** (4):462-469.

Gravett M. G., Thomas A., Schneider K. A., Reddy A. P., Dasari S., Jacob T., Lu X., Rodland M., Pereira L., Sadowsky D. W., Roberts C. T. Jr., Novy M. J., & Nagalla S. R. (2007). Proteomic analysis of cervical-vaginal fluid: identification of novel biomarkers for detection of intra-amniotic infection. *J Proteome Res* **6** (1):89-96.

Grenelee R. T., Hill-Harmon M. B., Murray T., & Thun M. (2001). Cancer Statistics, 2001. CA *Cancer J Clin* **51** (1):15-36.

Griffiths H. Can peptidomics provide a useful approacyh for the identification of biomarkers of toxicological exposure or effect? In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Grus F. H., Joachim S. C., & Pfeiffer N. (2007). Proteomics in ocular fluids. *Proteomics Clin Appl* **1** (8):876-888.

Horn D. M., Zubarev R. A., & McLafferty F. W. (2000). Automated *de novo* sequencing of proteins by tandem high-resolution mass spectrometry. *PNAS* 97 (19):10313-10317.

Hortin G.L. (2006). The MALDI-TOF Mass Spectrometric View of the Plasma Proteome and Peptidome. *Clin Chem* **52** (7):1223-1237.

Hoyert D. L., Mathews T. J., Menacker F., Stobino D. M., & Guyer B. (2006). Annual summary of vital statistics: 2004. *Pediatrics* **117** (1):168-183. [Erratum, *Pediatrics* (2006) **117** (6):2338]

Hu S., Loo J. A., Wong D. T. (2006). Human body fluid proteome analysis. *Proteomics* **6** (23):6326-6353.

Hughes J., Smith T. W., Kosterlitz H. W., Fothergill L. A., et al. (1975). Identification of two related pentapeptides from the brain with potent opiate agonist activity. *Nature* **258** (5536):577-580.

Iams J., Goldenberg R., Meis P., Mercer B., Moawad A., Das A., Thom E., McNellis D., Copper R., Johnson F., Roberts J., Miodovnik M., Van Dorsten J., Caritis S., Thurnau G., Bottoms S., and the NICHD Maternal Fetal Medicine Units Network. (1996). The length of the cervix and the risk of spontaneous premature delivery. *N Eng J Med* **334** (9):567-572.

Issaq H. J., Conrads T. P., Janini G. M., & Veenstra T. D. (2002). Methods for fractionation, separation and profiling of proteins and peptides. *Electrophoresis* **23** (17):3048-3061.

Jacobs J. M., Adkins J. N., Qian W. J., Liu T., Shen Y., Camp D. G. II, & Smith R. D. (2005). Utilizing human blood plasma for proteomic biomarker discovery. *J Proteome Res* **4** (4):1073-1085.

Johansson C., Samskog J., Sundstrom L., Wadensten H., Björkesten L., & Flensburg J. (2006). Differential expression analysis of Escherichia coli proteins using a novel software for relative quantitation of LC-MS/MS data. *Proteomics* **6** (16):4475-4485.

Jolley W. B., & Hinshaw D. B. (1965). Basic peptides isolated from thymus gland and blood and their possible role in the momograft reaction. *Surg Forum* **16**:211-213.

Kahn P. (1995). Molecular biology: from genome to proteome: looking at a cell's proteins. *Science* **270** (5235):369-370.

Kaiser T., Hermann A., Kielstein J. T., Wittke S., et al. (2003). Capillary electrophoresis coupled to mass spectrometry to establish polypeptide patterns in dialysis fluids. *J Chromatogr A* **1013** (1-2):157-171

Kendrew J. C., Bodo G., Dintzis H. M., Parrish R. G., Wyckoff H., & Phillips D. C. (1958). A Three-Dimensional Model of the Myoglobin Molecule Obtained by X-Ray Analysis. *Nature* **181** (4610): 662–666.

Kishor M., & Cable G. (2003). Meta-analysis of prostate-specific antigen and digital rectal examination as screening tests for prostate carcinoma. J *Am Board Fam Pract* **16** (2):95-101.

Koomen J. M., Zhao H., Li D., Nasser W., Hawke D. H., Abbruzzese J. L., Baggerly K. A., & Kobayashi R. (2005). Diagnostic protein discovery using liquid chromatography/mass spectrometry for proteolytic peptide targeting. *Rapid Commun Mass Spectrom* **19** (12):1624-1636.

Kozak K. R., Su F., Whitelegger J. P., Faull K., Reddy S., & Farias-Eisner R. (2005). Characterization of serum biomarkers for detection of early stage ovarian cancer. *Proteomics* **5** (17):4589-4596.

Kuriyama M., Wang M. C., Papsidaro L. D., et al. (1980). Quantitation of prostate-specific antigen in serum by a sensitive enzyme immunoassay. *Cancer Res* **40** (12):4658-62.

Laemmli U. K. (1970). Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227** (5259):680-685.

Le L., Chi K., Tyldesley S., Flibotte S., Diamond D. L., Kuzyk M. A., Sadar M. D. (2005). Identification of serum amyloid A as a biomarker to distinguish prostate cancer patients with bone lesions. *Clin Chem* **51** (4):695-707.

Li J., Zhang Z., Rosenzweig J., Wang Y. Y., Chan D. W. (2002). Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem* **48** (8):1296-1304.

Li J., Orlandi R., White C. N., Rosenzweig J., Zhao J., Seregni E., et al. (2005). Independent validation of candidate breast cancer serum biomarkers identified by mass spectrometry. Clin Chem **51** (12):2229-2235.

Liao H., Wu J., Kuhn E., Chin W., et al. (2004). Arthritis Rheum **50** (12):3792-3803.

Liotta L. A., Ferrari M., Petricoin E. (2003). Clinical proteomics: written in blood. *Nature* **425** (6961):905.

Loeb S., & Catalona W. J. (2007). Prostate-specific antigen in clinical practice. *Cancer Letters* **249** (1):30-39.

Loo J. A., Udseth H. R., Smith R. D. (1988). Collisional effects on the charge distribution of ions from large molecules, formed by electrospray-ionization mass spectrometry. *Rapid Commun Mass Spectrom* **2** (10):207-210.

Lucas F., Barber M., and Wolstenholme W. (1969). Mass-spectrometric determination of the amino acid sequences in peptides isolated from protein silk fibroin of *Bombyx mori.* Biochem J **114** (4):695-702.

Lundblad RL. The evolution from protein chemistry to proteomics. Taylor & Francis Group, Boca Raton, FL 295 **2006** pp. 1-12.

Malik G., Ward M. D., Gupta S. K., Trosset M. W., Grizzle W. E., Adam B. L., Diaz J. I., & Semmes O. J. (2005). Serum levels of an isoform of apolipoprotein A-II as a potential marker for prostate cancer. *Clin Cancer Res* **11** (3):1073-1085.

Marshall J., Kupchak P., Zhu W., Yantha J., Vrees T., Furesz S., et al. (2003). Processing of serum proteins underlies the mass spectral fingerprinting of myocardial infarction. *J Proteome Res* **2** (4):361-372.

McLafferty F. W. (2001). Tandem mass spectrometric analysis of complex biological mixtures. *Int J Mass Spectrom* **212** (1-3):81-87.

Meng Z., & Veenstra T. D. (2007). Proteomic analysis of serum, plasma, and lymph for the identification of biomarkers. *Proteomics Clin Appl* **1** (8):747-757.

Merrell K., Southwick K., Graves S. W., Esplin M. S., Lewis N. E., & Thulin C. D. (2004). Analysis of low-abundance, low-molecular-weight serum proteins using mass spectrometry. *J Biomol Tech* **15** (4):238-248.

Merrell K., Thulin C. D., Esplin M. S., & Graves S. W. (2008). Systematic internal standard selection for capillary liquid chromatography-mass spectrometry time normalization to facilitate serum proteomics. *J Biomol Tech* **19** (5):320-327.

Mikolajczyk, S. D., Song, Y., Wong, J. R., Matson, R. S., & Rittenhouse, H. G. (2004). Are multiple markers the future of prostate cancer diagnostics? *Clin Biochem* **37** (7), 519-528.

Mischak H., Julian B. A., & Novak J. (2007). High-resolution proteome/peptidome analysis of peptides and low-molecular-weight proteins in urine. *Proteomics Clin Appl* **1** (8):792-804.

Mitchell B. L., Yasui Y., Lampe J. W., Gafken P. R., & Lampe P. D. (2005). Evaluation of matrix-assisted laser desorption/ionization-time of flight mass spectrometry proteomic profiling: identification of α2-HS glycoprotein B-chain as a biomarker of diet. *Proteomics* **5** (8):2238-2246.

Mosesson M. W. (2005) Fibrinogen and fibrin structure and functions. *J Thromb Haemost* **3** (8):1894-1904.

Mujezinovic N., Raidl G., Hutchins J. R. A., Peters J. M., Mechtler K., & Eisenhaber F. (2006). Cleaning of raw peptide MS/MS spectra: Improved protein identification following deconvolution of multiply charged peaks, isotope clusters, and removal of background noise. *Proteomics* **6** (19):5117-5131.

Nomura F., Tomonaga T., Sogawa K., Ohashi T., Nezu M., Sunaga M., Kondo N., Iyo M., Shimada H., & Ochiai T. (2004). Identification of novel and downregulated biomarkers for alcoholism by surface enhanced laser desorption/ionization-mass spectrometry. *Proteomics* **4** (4):1187-1194.

O'Farrell P. H. (1975). High resolution two-dimensional electrophoresis of proteins. *J Biol Chem* **250** (10):4007-4021.

Omenn G. S., States D. J., Adamski M., Blackwell T. W., Menon R., Hermjakob H., et al. (2005). Overview of the HUPO Plasma Proteome Project: results from the pilot phase with 35 collaborating laboratories and multiple analytical groups, generating a core dataset of 302 proteins and a publicly-available database. *Proteomics* **5** (13):3226-3245.

Papayannopoulos I. A. (1995). The interpretation of collision-induced dissociation tandem mass spectra of peptides. *Mass Spectrom Rev* **14** (1):49-73.

Patterson SD, Abersold R, Goodlett DR. In: Pennington SR, Dunn MJ, editors. Proteomics from protein sequence to function. BIOS Scientific Publishers Ltd. 2001.

Pereira L., Reddy A. P., Jacob T., Thomas A., Schneider K. A., Dasari S., Lapidus J. A., Lu X., Rodland M., Roberts C. T. Jr., Gravett M. G., & Nagalla S. R. (2007). Identification of novel protein biomarkers of preterm birth in human cervical-vaginal fluid. *J Proteome Res* **6** (4):1269-1276.

Petricoin E. F., Ardekani A. M., Hitt B. A., Levine P. J., Fusaro V. A., Steinberg S. M., Mills G. B., Simone C., Fishman D. A., Kohn E. C., & Liotta L. A. (2002). Use of proteomic patterns in serum to identify ovarian cancer. *Lancet* **359** (9306):572-577.

Petricoin E. F., Paweletz C. P., Liotta L. A. (2002). Clinical applications of proteomics: Proteomic pattern diagnostics. *J Mammary Gland Biol & Neoplasia* **7** (4):433-440.

Petricoin E. F., & Liotta L. A. (2003). Counterpoint--The vision for a new diagnostic paradigm. *Clin Chem* **49** (8), 1276-1278.

Petricoin E. F., & Liotta L. A. (2003). Clinical applications of proteomics. *J Nutr* **133** (7):2476S-2484S.

Petricoin E. F., Liotta L. A. (2006). A revolutionary approach to biomarker discovery. *Scientist* **20** (11):32-39.

Piñeiro M., Andrés M., Iturralde M., Carmona S., Hirvonen J., Pyörälä S., Heegaard P. M., Tjørnehøj K., Lampreave F., Piñeiro A., & Alava M. A. (2004). ITIH4 (*inter-alpha-trypsin inhibitor heavy chain 4*) is a new acute-phase protein isolated from cattle during experimental infection. *Infect Immun* **72** (7):3777-82.

Qian W. J., Jacobs J. M., Liu T., Camp D. G. II, & Smith R. D. (2006). Advances and challenges in liquid chromatography-mass spectrometry-based proteomics profiling for clinical applications. *Mol Cell Proteomics* **5** (10)*,*1727-1744.

Radulovic D., Jelveh S., Ryu S., Hamilton T.G., et al. (2004). *Mol Cell Proteomics* **3** (10):984-997.

Richter R., Schulz-Knappe P., Schrader M., Ständker L., Jurgens M., Tammen H., & Forssmann W. G. (1999). Composition of the peptide fraction in human blood plasma: database of circulating human peptides. *J Chromatogr B* **726** (1-2):25-35.

Ruetschi U., Rosen A., Karlsson G., Zetterberg H., Rymo L., Hagberg H., & Jacobsson B. (2005). Proteomic analysis using protein chips to detect biomarkers in cervical and amniotic fluid in women with intra-amniotic inflammation. *J Proteome Res* **4** (6):2236-2242.

Ruhlen R. L., & Sauter E. R. (2007). Proteomics of nipple aspirate fluid, breast cyst fluid, milk, and colostrum. *Proteomics Clin Appl* **1** (8):845-852.

Rush R. W., Keirse M. J., Howat P., Baum J. D., Anderson A. B., & Turnbull A. C. (1976). Contribution of preterm delivery to perinatal mortality. *Br Med J* **2** (6042):965-968.

Schrader M., & Schulz-Knappe P. (2001). Peptidomics technologies for human body fluids. *Trends Biotechnol* **19** (10 Suppl):S55-S60

Schulz-Knappe P., Schrader M., & Zucht H. D. (2005). Peptide sequence prediction supported by correlation-associated networks in human cerebrospinal fluid. *Comb Chem High Throughput Screen* **8** (8):697-704.

Semmes O. J., Cazares L. H., Ward M. D., Qi L., Moody M., Maloney, et al. (2005). Discrete serum protein signatures discriminate between human retrovirus-associated hematologic and neurologic disease. *Leukemia* **19** (7):1229-1238.

Shaw C, Verhaert PDEM. Peptidomics and biology: two scientific disciplines driving each other. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Sköld K., Svensson M., Kaplan A., Björkesten L., Aström J., & Andrén P. E. (2002). A neuroproteomic approach to targeting neuropeptides in the brain. *Proteomics* **2** (4):447-454.

Sköld K, Fälth M, Svensson M, Nilsson A, Svenningsson P, Andrén P. Strategies for reliable and improved identification of peptides. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Song J., Patel M., Rosenzweig C. N., Chan-Li Y., Sokoll L. J., Fung E. T., Choi-Miura N. H., Goggins M., Chan D. W., & Zhang Z. (2006). Quantification of fragments of human serum inter-alpha-trypsin inhibitor heavy chain 4 by a surface-enhanced laser desorption/ionization-based immunoassay. *Clin Chem* **52** (6):1045-53.

Srinivas P. R., Kramer B. S., Srivastava S. (2001). Trends in biomarker research for cancer detection. *Lancet Oncol* **2** (11):698-704.

Strand F. L. (2003). Neuropeptides: general characteristics and neuropharmaceutical potential in treating CNS disorders. *Prog Drug Res* 61:1-37.

Sunderland T., Hampel H., Takeda M., Putnam K. T., Cohen R. M. (2006). Biomarkers in the diagnosis of Alzheimer's disease: are we ready? *J Geriatr Psychiatry Neurol* **19** (3):172-179.

Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Seer*Stat Database: Incidence – SEER 17 Regs Public-use, Nov 2005 Sub (1973-2005 varying), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2006, based on the November 2005 submission. Accessed December 7th, 2008.

Svensson M., Skold K., Svenningson P., & Andren P. E. (2003). Peptidomics-based discovery of novel neuropeptides. *J Proteome Res* **2** (2):213-219.

Thadikkaran L., Siegenthaler M. A., Crettaz D., Queloz P. A., Schneider P., & Tissot J. D. (2005). Recent advances in blood-related proteomics. *Proteomics* **5** (12):3019-3034.

Tirumalai R. S., Chan K. C., Prieto D. A., Issaq H. J., Conrads T. P., & Veenstra T. D. (2003). Characterization of the low molecular weight human serum proteome. *Mol Cell Proteomics* **2** (10): 1096-1103.

Tolson J., Bogumil R., Brunst E., Beck H., Elsner R., Humeny A., et al. (2004). Serum protein profiling by SELDI mass spectrometry: detection of multiple variants of serum amyloid αin renal cancer patients. *Lab Invest* **84** (7):845-856.

Torikai E., Kaegyama Y., Takahashi M., & Nagano A. (2006). The effect of methotrexate on bone metabolism markers in patients with rheumatoid arthritis. *Mod Rehumatol* **16** (6):350-354.

Torikai E., Kageyama Y., Takahashi M., Suzuki M., Ichikawa T., Nagafusa T., & Nagano A. (2006). The effect of infliximab on bone metabolism markers in patients with rheumatoid arthritis. *Rheumatology* **45** (6):761-764.

Vermeulen R., Lan Q., Zhang L., Gunn L., McCarthy D., Woodbury R.L., et al. (2005). Decreased levels of CXC-chemokines in serum of benzene-exposed workers identified by array-based proteomics. *Proc Natl Acad Sci* **102** (47):17041-17046.

Villaneuva J., Shaffer D. R., Philip J., Chaparro C. A., Erdjument-Bromage H., Olshen A. B., et al. (2006). Differential exopeptidase activities confer tumor-specific serum peptidome patterns. *J Clin Invest* **116** (1):271-284.

Wasinger V., Cordwell S., Cerpa-Poljak A., Yan J., Gooley A., Wilkins M., Duncan M., Williams K., & Humphrey-Smith I. (1995). Progress with gene-product mapping of the Mollcules: *Mycoplasma genitalium*. *Electrophoresis* **16** (1):1090-1094.

Wittke S., Fliser D., Haubitz M., Barte S., et al. (2003). Capillary electrophoresis coupled to mass spectrometry to establish polypeptide patterns in dialysis fluids. *J Chromatog A* **1013** (1-2):173-181.

Woong-Shick A., Sung-Pil P., Su-Mi B., Joon-Mo L., Sung-Eun N., Gye-Hyun N., et al. (2005). Identification of hemoglobin-α and -ß subunits as potential serum biomarkers for the diagnosis and prognosis of ovarian cancer. *Cancer Sci* **96** (3):197-201.

Wu K., & Zhang Y. (2007). Clinical application of tear proteomics: Present and future prospects. *Proteomics Clin Appl* **1** (9):972-982.

Wulfkuhle J. D., Paweletx C. P., Steeg P. S., Petricoin E. F., & Liotta L. (2003). Proteomic approaches to the diagnosis, treatment, and monitoring of cancer. *Adv Exp Med Biol* **532**:59-68.

Yim E. K., Park J. S. (2006). Role of proteomics in translational research in cervical cancer. *Expert Rev Proteomics* 3 (1):21-36.

Zhang Z., Bast R. C. Jr., Yu Y., Li J., Sokoll L. J., Rai A. J., et al. (2004). Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer. *Cancer Res* **64** (16):5882-5890.

Zhang H., Liu A. Y., Loriaux P., Wollscheid B., Zhou Y., Watts J. D., & Abersold R. (2007). Mass spectrometric detection of tissue proteins in plasma. *Mol Cell Proteomics* **6** (1):64-71.

Zolotarjova N., Martosella J., Nicol G., Bailey J., et al. (2005). Differences among techniques for high-abundant protein depletion. *Proteomics* **5** (13):3304-3313.

Zubarev RA, Zubarev AR, Savitski MM. Electron capture/transfer versus collisionally activated/induced dissociations; solo or duet? J Am Chem Soc 2008;19:753-761.

Zürbig P., Renfrow M. B., Schiffer E., Novak J., et al. (2006). *Electrophoresis* **27** (11):2111-2125.

Zubrig P, Mischak H. Peptidomics approach to proteomics. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).

Zübrig P, Mischak H. Biomarker Discovery. In: Soloviev M, Shaw C, Andrén P, editors Peptidomics Methods and Applications. John Wiley & Sons, Inc., Hoboken, New Jersey (2008).