BIOMARKER DISCOVERY IN AUTOIMMUNE DISEASES:

A PROTEOMICS APPROACH

APPROVED BY SUPERVISORY COMMITTEE

Thomas Kodadek, Ph.D. (Mentor)
Harold Garner, Ph.D. (Chair)
Peter Antich, Ph.D.
Marc Mumby, Ph.D.
Kevin Rosenblatt, M.D. Ph.D.

ACKNOWLEDGMENTS

I would like to extend my utmost gratitude to my mentor, Prof. Tom Kodadek, for being a constant source of support, advice, guidance and encouragement. He is not only a brilliant scientist but also an outstanding mentor who truly cares about his students. I am thankful to him for giving this electrical engineering student a shot at the life sciences in his laboratory and for introducing me to the fascinating world of mass spectrometry and proteomics. I am forever indebted to Tom for the invaluable training he has given me in becoming an independent scientist with an analytical mind and for providing a friendly, stimulating and conducive lab environment for research.

I am also most appreciative to my committee members, Drs. Harold 'Skip' Garner, Kevin Rosenblatt, Marc Mumby and Peter Antich, for their constructive criticisms and for always making the time for me and my projects. My collaborators, Skip and Kevin, have been more than willing in extending a helping hand to me over the years, be it in administrative issues or in research, and are without a doubt instrumental in the successful completion of my projects. I thank them for their sage advice and guidance and for granting me unrestricted access to their labs' resources, especially to the talents of Prem Gurnani and Wayne Fisher.

I have been very blessed to be in the accompaniment of the gifted and friendly members of the Kodadek Lab in particular, and of the CBI/DTR as a whole. I have learnt a lot from each and every one of them over the years, both in science and life lessons. I am particularly grateful for my partners in crime, Reddy Moola, Sara Chirayil, Danielle

Miller and Hyun-suk Lim, for keeping my sanity through words of encouragement in times of despair and for sharing my moments of joy. I thank them for their sincere friendship and intriguing conversations over cake every Friday.

I am also appreciative to Dr. Bridgette Kirkpatrick with whom I took my first few biology classes during my transition from engineering to the life sciences and who encouraged me to pursue a career in the sciences. Even though my research skills got perfected here in Dallas, it really began in Plano with her.

Last but not least, I would like to thank my family members and friends for their constant support and encouragement. In particular, my sister Chinny and brother-in-law Son have been supportive in everything that I do in life, not just in my Ph.D. career. To all my friends on and off campus, thank you for all the happy hours and for making my life in Texas bearable.

BIOMARKER DISCOVERY IN AUTOIMMUNE DISEASES:

A PROTEOMICS APPROACH

by

CHIANG "NICLAS" TAN

DISSERTATION

Presented to the Faculty of the Graduate School of Biomedical Sciences

The University of Texas Southwestern Medical Center at Dallas

In Partial Fulfillment of the Requirements

For the Degree of

DOCTOR OF PHILOSOPHY
The University of Texas Southwestern Medical Center at Dallas

Dallas, Texas

November, 2008

BIOMARKER DISCOVERY IN AUTOIMMUNE DISEASES:

A PROTEOMICS APPROACH

CHIANG "NICLAS" TAN, Ph.D.

The University of Texas Southwestern Medical Center at Dallas, 2008

THOMAS KODADEK, Ph.D.

Proteins constitute the functional machinery of cells and are prime candidates for disease marker discovery. Mass spectrometry-based proteomics biomarker discovery holds the ability to interrogate a constellation of proteins simultaneously in a high-throughput manner to uncover a panel of markers that are specific to the presence of a disease. However, the rate of introduction of novel biomarkers with clinical currency has declined in the past few years due to challenges faced by both the discovery and validation stages. Surface retentate chemistry-mass spectrometry is a powerful platform that allows on-chip simplification of complex biological samples to better match the current dynamic range

of mass spectrometers. Initial reports on differential protein profiling using this approach produced profiles with high sensitivities and specificities for disease classification. As with any maturing technology, issues that were overlooked during its introduction are now the main barriers to its clinical utility. The improved workflow described here aims to address some of these pending issues. Specifically, the experimental design incorporated knowledge of the disease pathway into sample selection, elected sample sources that are rich in diagnostic markers, and adopted biological and technical replicates to minimize variance. To ensure reproducibility, complete automation of the process from sample preparation to data acquisition was incorporated along with the adoption of a high performance mass spectrometer with minimal mass drift. A robust data analysis approach was implemented to overcome the issue of overfitting and to effectively trim down the list of candidate biomarkers to the selected few with true discriminatory power to facilitate downstream validation. As a demonstration of the robustness and utility of the workflow, profiling studies were performed on two autoimmune diseases. Protein profiles with high mass peak fidelity were obtained with high discriminatory power. Selective differential peaks were further investigated and confirmed to display differential levels in clinical samples. Validation in a larger sample set should determine the diagnostic potential of these markers for clinical application. Finally, a high-throughput study is reported showing that peptoids are, in general, a relatively more cell permeable class of molecules than peptides, rendering them ideal for drug development to target disease biomarkers.

TABLE OF CONTENTS

PRIOR PUBLICATIONS

(1) **Tan, N.C.**, Gurnani, P., Fisher, W., Krastins, B., Sarracino, D., Lopez, M.F., Garner, H., Rosenblatt, K.P., and Kodadek, T., "Proteomics Biomarker Discovery in Narcolepsy: A High-Throughput, Blood-Based Pilot Study". *Manuscript in preparation*.

(2) **Tan, N.C**., Gurnani, P., Fisher, W., Racke, M.K., Lovett-Racke, A.E., Garner, H., Rosenblatt, K.P., and Kodadek, T., "Proteomics Biomarker Discovery in Multiple Sclerosis: Complement C3 as a Cerebrospinal Fluid Marker for Disease Progression". *Submitted*.

(3) **Tan, N.C**., Fisher, W., Rosenblatt, K.P., and Garner, H., "Application of Multiple Statistical Tests to Enhance Mass Spectrometry-Based Biomarker Discovery". *Submitted*.

(4) **Tan, N.C.**, Yu, P., Kwon, Y., and Kodadek, T., (2008) "High-Throughput Evaluation of Relative Cell Permeability between Peptoids and Peptides". *Bioorg Med Chem* 16(11), 5853-61.

(5) Fisher, W.G., Rosenblatt, K.P., Fishman, D.A., Whiteley, G.R., Mikulskis, A., Kuzdzal, S.A., Lopez, M.F., **Tan, N.C.**, German, D.C., Garner, H.R., (2007) "A Robust Biomarker Discovery Pipeline for High-Performance Mass Spectrometry Data". *J Bioinform Comput Biol* 5(5), 1023-45.

LIST OF FIGURES

LIST OF TABLES

LIST OF APPENDICES

ABBREVIATIONS

2DGE:	Two-dimensional gel electrophoresis
AIC:	Akaike information criterion
AUC:	Area under ROC curve
BBB:	Blood-brain barrier
CART:	Classification and regression tree
CHCA:	α-cyano-4-hydroxycinnamic acid
CID:	Collision induced dissociation
ClogP:	Calculated n-octanol–water partition coefficient
CNS:	Central nervous system
CSF:	Cerebrospinal fluid
CV:	Coefficient of variation
EAE:	Experimental autoimmune encephalomyelitis
EDS:	Excessive daytime sleepiness
ELISA:	Enzyme-linked immunosorbent assay
ESI:	Electrospray ionization
FDA:	Food and Drug Administration
FT-ICR:	Fourier transform ion cyclotron resonance
Gof:	Goodness of fit
HLA:	Human leukocyte antigen
HMW:	High molecular weight
HUPO:	Human Proteome Organization
LC:	Liquid chromatography

LMW: Low molecular weight

IMAC: Immobilized metal affinity chromatography

MALDI: Matrix-assisted laser desorption/ionization

MHC: Major histocompatibility complex

MRM: Multiple reaction monitoring

MS: Mass spectrometry

MS/MS: Tandem mass spectrometry

MSLT: Multiple sleep latency test

MudPIT: Multidimensional protein identification technology

m/z: mass-to-charge ratio

NIST: National Institute of Standards and Technology

NPV: Negative predictive value

OND: Other neurological diseases

OxDex: Dex-17β-carboxylic acid

PD: Parkinson's disease

PPMS: Primary progressive multiple sclerosis

PPV: Positive predictive value

PR: Permeability ratio

PSA: Prostate specific antigen

PTM: Post-translational modification

RA: Rheumatoid arthritis

REM: Rapid eye movement

ROC: Receiver operating characteristics

RRMS: Relapsing-remitting multiple sclerosis

SA: Sinapinic acid

SDex: Dex-21-thiopropionic acid

SELDI: Surface-enhanced laser desorption/ionization

SISCAPA: Stable isotope standard and captured anti-peptide antibody

SPMS: Secondary progressive multiple sclerosis

TIC: Total ion current

TOF: Time-of-flight

TPSA: Topological polar surface area

TTR: Transthyretin

UPGMA: Unweighted pair group method with arithmetic mean

CHAPTER ONE
Introduction

1.1 THE BIRTH OF PROTEOMICS

The Human Genome Project was officially completed in 2003 [1, 2]. The approximately three billion nucleotides sequenced contain within them a trove of cryptic information that ultimately dictates the synthesis of diverse molecules that govern the human body. A surprising discovery from this massive, international endeavor was that the human genome is represented by ca. 30,000 genes, which is only twice as many as less sophisticated organisms such as the worm or the fly [2]. It was further revealed that only between 21,000 and 23,000 are protein-coding genes [3, 4] of which a meager 10,000 are active at any given time to reflect the current physiological state of the cell. This reinforces the notion that cellular processes are built up by complex networks of specific interactions of protein molecules whose diversity, structural and functional information cannot be ascertained from the genomic sequences alone.

Consequently, attention was shifted onto the proteome, which by definition is a dynamic collection of proteins encoded by the genome that demonstrate variation between individuals, between cell types, and between entities of the same type but under different pathological and physiological conditions [5, 6]. The study of the proteome spawned the new field of proteomics, which now evokes not only the identification and quantification of all the proteins in any given cell, but also the set of all protein isoforms and modifications, the interactions between them, their localization, turnover, activities and function [7].

The sudden heightened interest in proteomics was propelled by two main concurrent scientific advances: the availability of gene and genomic sequence databases, and technological advances in protein analytical tools. Even though the study of protein structure and function has been the focus of biochemical research for years, determining the identity of proteins was difficult because of a lack of sensitive and rapid analytical methods for protein characterization (such as the attendant polymerase chain reaction and automated sequencing technologies that are

readily available for DNA analysis in genomics). These limitations were overcome by the emergence of mass spectrometry with soft ionization techniques such as matrix-assisted laser desorption/ionization (MALDI) introduced by Tanaka *et al*. in 1988 [8, 9] and electrospray ionization (ESI) introduced by Fenn *et al*. [10] at around the same time. Both of these revolutionary technologies garnered the Nobel Prize in Chemistry in 2002. These techniques allow peptide and protein structures to be retained during the ionization process without inducing fragmentation, and when coupled with the availability of the entire human coding sequence in genome databases, enable the identification of proteins and determination of their primary structure in a high-throughput, rapid and facile manner [11]. Proteomics was further facilitated in large-scale protein analyses by the subsequent seeding of the auxiliary field of bioinformatics where elaborate computational tools were conceived to store, process, analyze and interpret the large amounts of data generated.

To date, mass spectrometry-based proteomics encompass three main areas of research: (i) protein identification initiatives to annotate all proteins within a biological specimen as coordinated by the Human Proteome Organization (HUPO), (ii) biomarker search in body fluids, cells and tissues for disease diagnosis, prognosis and response to therapeutic treatments and (iii) interrogation of protein-protein interactions to identify and characterize protein-binding partners. This dissertation will only focus on the application of mass spectrometry in biomarker discovery.

1.2 PROTEIN BIOMARKERS

1.2.1 The case for proteins

The highly acclaimed genomic blueprint of our species made available by the Genome Project has yet to fulfill its promise of revolutionizing biology and medicine by unraveling causal relationships between genes and disease onset. Indeed, the baton has now been passed on to proteomics. In retrospect, this is hardly surprising as genetic composition is a static state which at best predicts the participation of proteins in a cell, whereas cell operation is a dynamic process orchestrated by expressed proteins. Proteins are the functional cellular entities that partake in almost all the biochemical activities in the cell, ranging from transcription factors and enzymes involved in cellular pathways to antibodies and cytokines that dominate the immune response.

A protein's function is dependent on its structure and complex interactions with other biomolecules, none of which can be predicted accurately from the sequence information alone. Studies have shown that there is a poor correlation (<0.5) between mRNA and protein expression levels [12-14]. This can be attributed to a myriad of post-transcriptional regulatory mechanisms. Moreover, a single gene can encode multiple different proteins through the process of alternative splicing of primary transcripts where multiple mature transcripts can be obtained from the same gene resulting in the translation of related but different proteins (alternative splice variants), and from the presence of sequence polymorphisms. This disjunction between mRNA and protein levels is further amplified by the greater than 200 post-translational modifications (PTMs) [15] that could embellish any protein in the cell [13, 16].

Furthermore, what makes biomarker discovery at the protein level so appealing is that the onset of diseases causes the protein composition of the system to change to reflect its new disease state. This is illustrated in Figure 1.1 which depicts the general time course of development of autoimmune diseases, such as narcolepsy, Type 1 diabetes and multiple sclerosis. Autoimmune

disease patients are usually genetically predisposed and are positive for susceptibility genes from the human leukocyte antigen or HLA region. However, being positive for these genes does not always cause the disease to develop. The precipitation of the disease is usually brought about by environmental factors. In the case of neurodegenerative diseases, upon exposure to these environmental factors, the neurons begin to sustain minor injury. With time, the number of neurons affected will reach a threshold level where clinical symptoms begin to appear. The pathogenesis of these diseases usually begins 5 to 10 years before early symptoms are presented. Therefore, direct analysis at the protein level provides a more encompassing view of critical changes that occur at the molecular level due to a disease. Proteomics not only offers the capability to confirm the presence of these disease-related proteins but also provides a direct measure of their abundance.

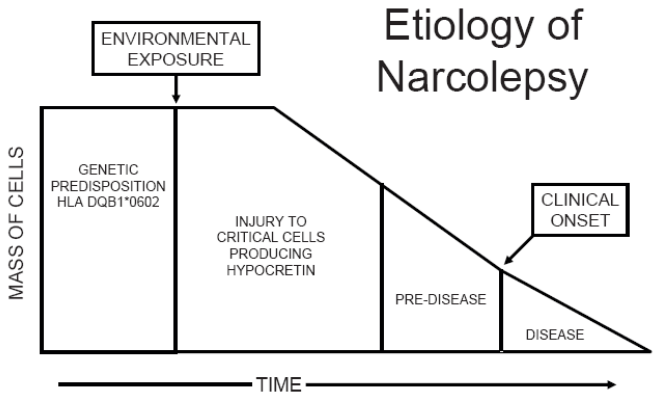


Figure 1.1 **Model depicting development of narcolepsy**. Disease onset usually begins at the molecular level long before presentation of symptoms. [17]

1.2.2 Protein biomarkers as diagnostic entities

A biomarker is defined as a characteristic that is objectively measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic response to a therapeutic intervention [18]. Biomarkers can come from many places, including transcriptional profiling and DNA methylation studies in cancer [19] and metabolomics profiling studies in metabolic diseases [20]. Genetic biomarkers do not change with time or age and can be good predictors of disease susceptibility risk and responsiveness to therapeutic regimes. Unfortunately, they are not suitable for screening or confirmatory diagnosis as being positive for a susceptibility gene does not guarantee disease precipitation. As mentioned above, environmental factors are culpable for disease onset in complex diseases such as cancer and autoimmune conditions and their influence is reflected in the perturbation of the cell's proteome composition. As such, the proteome is likely the most ubiquitously affected in disease and for this very reason, protein biomarkers are actively sought to complement genetic biomarkers. Proteomics can contribute directly to biomarker and drug development as almost all drugs are directed against proteins with the major exception of genotoxic anti-cancer drugs. Furthermore, the proteomic measurement already delivers the desired end point, namely the protein expression level of the gene of interest. The search for biomarkers in proteins has an established history in clinical chemistry, with precedences set by the likes of prostate-specific antigen (PSA) for the diagnosis of prostate cancer, Bence Jones Protein in urine for multiple myeloma, and CA-125 for ovarian cancer.

There has been a surge of enthusiasm surrounding molecular diagnostics due to its array of applications in disease management. Biomarkers are most commonly utilized for diagnosis of disease in established cases (e.g. elevated blood glucose concentration for the diagnosis of diabetes mellitus). Biomarkers are also a role player in the early detection of disease. Diagnosis in the asymptomatic early stages coupled with early intervention dramatically improves survival rates as seen in cancer [21]. Novel disease markers can contribute to this as an add on to the

current armamentarium of diagnostic tools by presenting itself as a primary non-invasive test before a more involved procedure is performed, such as measuring PSA in blood for early diagnosis prior to confirmatory biopsies. In addition to early detection, molecular markers can also serve as a tool for staging of disease, in the hope that the quality of life of the patients would be dramatically improved when timely interventions could be introduced at a more treatable stage of a disease. Novel protein biomarkers may also serve as great candidate therapeutic targets for drug development [22].

Protein molecular markers are assuming a prominent role in the prediction of drug efficacy with the dawn of personalized medicine. This is exemplified by the drug Herceptin (Genentech) which is only effective in breast cancer patients who express the HER2 biomarker, and Gleevec (Novartis) which is ineffective against chronic myeloid leukemia patients who have a mutation in the marker BCR-ABL. Due to their indispensable value for monitoring drug efficacy, treatment selection and dosage determination, it may be plausible for the US Food and Drug Administration (FDA) to require a biomarker to accompany each drug that is destined for the consumer market in the near future.

In spite of the importance of protein markers, the rate of introduction of novel biomarkers into clinical practice is extremely disappointing [23, 24]. Indeed, since 1998, the rate of introduction of new protein analytes approved by the FDA has fallen to an average of one per year [25]. In cancer screening, only a handful of markers have become widely accepted by the clinical community, including cancer antigens 15.3, 19.9 and 125, carcinogenic embryonic antigen, PSA, and human papillomavirus, to name a few [26]. These scarce biomarkers themselves suffer from low sensitivity and specificity. For example, CA-125 which was adopted as a marker for ovarian cancer, is only effective in the advanced stages but not early stages, with a positive predictive value (PPV) of only 10% due to the fact that CA-125 is also expressed in other diseases [27]. Serum PSA has been adopted as a monitoring test for prostate cancer since the 1980s, successfully impacting disease management and monitoring in prostate cancer from

greatly reduced presentation of men with advanced stages of the disease. However, its purported high sensitivity is shadowed by its low specificity of only 20 to 40% and failure to detect the majority of prostate cancers, including those that are high grade and those with PSA levels below 4 ng/ml [28]. Because PSA is prostate specific, and not prostate cancer specific, increased concentrations of PSA are also found in benign prostatic hyperplasia, acute and chronic prostatitis, and prostatic intraepithelial neoplasia [29].

There is an urgent need for biomarkers that are disease specific to improve diagnosis, guide molecularly targeted therapy and monitor therapeutic response across a wide spectrum of disease. For heterogeneous diseases like cancer and autoimmune conditions, it is probably impossible to unearth a single ‘magic bullet’ marker that will detect all subtypes with high specificity and sensitivity. However, both sensitivity and specificity can be attained via a panel of biomarkers. There is growing consensus that multiple markers will be required for most diagnostic applications in the future. A pattern of multiple biomarkers will undoubtedly contain a higher level of discriminatory information compared to a single biomarker alone, particularly for large heterogeneous patient populations, as confirmed in several recent publications [30, 31].

1.3 PROTEOMICS BIOMARKER DISCOVERY

Comparative proteomics showcases a platform that is suitable for the identification and even the verification of the novel biomarker panels. The hypothesis is that the proteome of biological samples from disease patients will vary dramatically from those of healthy, control patients and that this difference is attributable to the differential proteins that are either directly related to the disease pathogenesis or a consequence of it. Therefore, semiquantitative comparisons of relative protein abundance between disease and control patient samples can be used to identify proteins that are differentially present [32-34] to populate lists of potential biomarkers. This determination of the relative or absolute concentration of these molecules across sizeable number of specimens represents a key step toward providing insight into the physiological significance or diagnostic potential of the individual proteins. The era of unconventional discovery-based research, in lieu of its hypothesis-driven counterpart, has officially arrived, made feasible by new technologies that allow the proteome to be measured in greater detail and with increase speed.

De novo proteomics biomarker discovery demands a platform that is capable of detecting and quantifying protein marker candidates present at or below ng/ml levels in blood, where many disease-specific markers with clinical currency are thought to reside. PSA, for example, exists in the low ng/ml concentration in serum. Capture agents such as antibodies confer the sensitivity required and have been used in protein profiling array studies [35, 36]. This approach, however, is limited by the modest number of antibodies of sufficient quality that are currently available. The limitations of affinity approaches essentially leave mass spectrometry (MS) as the principal enabling technology for unbiased candidate protein marker discovery.

MS-based proteomics biomarker discovery is poised to address the paucity of biomarkers due to its ability to interrogate a constellation of proteins simultaneously in a high-throughput manner. In fact, recent biomarker discovery studies using unbiased approaches that couple high performance mass spectrometers and extensive sample processing have been fruitful in the

detection of low abundance proteins [37, 38], making MS the current preferred strategy for discovery of diagnostic, prognostic, and therapeutic protein biomarkers. Indeed, MS has the potential to revolutionize diagnostics by facilitating biomarker discovery, generating proteomic profiles as disease signatures [39], enabling tissue imaging [40] and quantifying biomarker levels [41]. However, as discussed below, this considerable potential has yet to be realized for a variety of reasons.

1.3.1 Challenges in proteomics biomarker discovery

Proteomics biomarker discovery is still in its infancy and faces numerous biological and technical hurdles that have to be overcome before it can deliver its promise of populating the list of candidate protein markers. These concerns include, but are not limited to, the enormous range of protein concentrations in complex samples such as blood, the limited dynamic range of current proteomic technologies, and the statistical challenges inherent in high-dimensionality data sets populated by comparatively few samples.

1.3.1.1 Proteome complexity

By far, blood is the biological sample of choice for numerous reasons: (i) blood is an established biological sample used in clinical diagnosis with assays to measure >100 proteins already in existence [42], (ii) it can be procured relatively non-invasively and is easy to handle, (iii) it is rich in proteins (estimated to contain tens of thousands of core proteins) at an average concentration of 80 mg/ml, and (iv) a majority of the protein constituents of the body can be found in blood. It is a circulating representation of all body tissues and of both physiological and pathological processes [23]. A multiplex of disease-specific analytes may be detectable in the blood, leading to convenience of testing [43].

The complexity of the blood proteome is a double-edged sword. The dynamic range of protein abundance spans ten to eleven orders of magnitude [23] (Figs. 1.2 and 1.3). Albumin itself exists at a normal concentration of 50 mg/ml in blood, overshadowing the low abundance, important disease biomarkers (such as PSA and Interleukin-6) that are usually present in the relatively low concentrations of ng and pg/ml due to massive dilution post-leakage into peripheral blood from diseased tissues. Indeed, the ten or so high abundance polypeptides that dominate the human plasma originate from a few major tissues, corroborating the search for diagnostic markers in the low abundance population if disorders of other tissues are of interest. Since low abundance proteins represent some of the more functionally important gene products, such as inflammatory molecules, transcription factors, and other regulatory proteins, whose aberrant expression contributes to disease onset, overcoming this dynamic range is essential to increase detection sensitivity. Cell line homogenates, tissue lysates and alternative biological fluids, such as urine and cerebrospinal fluid, which are also amenable for discovery efforts, also pose the dynamic range challenge, albeit to a lesser extent.

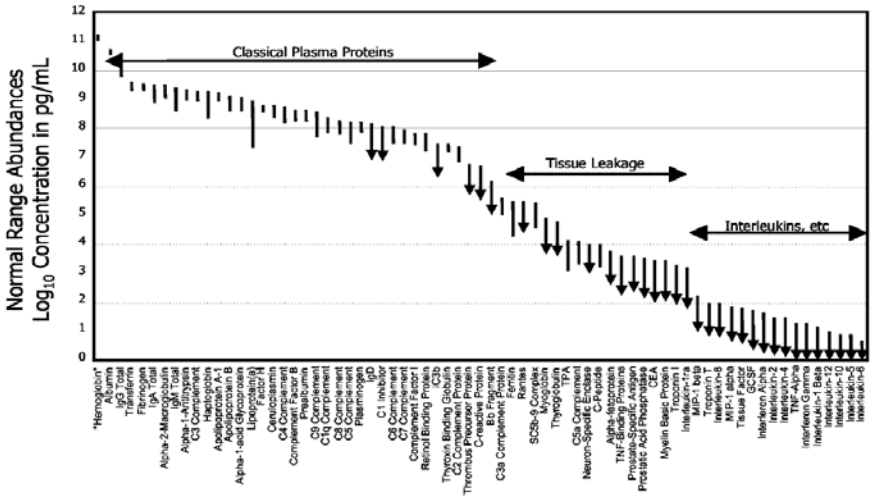


Figure 1.2 **Relative abundance of proteins in human plasma**. Abundance is plotted on a log scale spanning 12 orders of magnitude. Hemoglobin is included (*far left*) for comparison. [23]

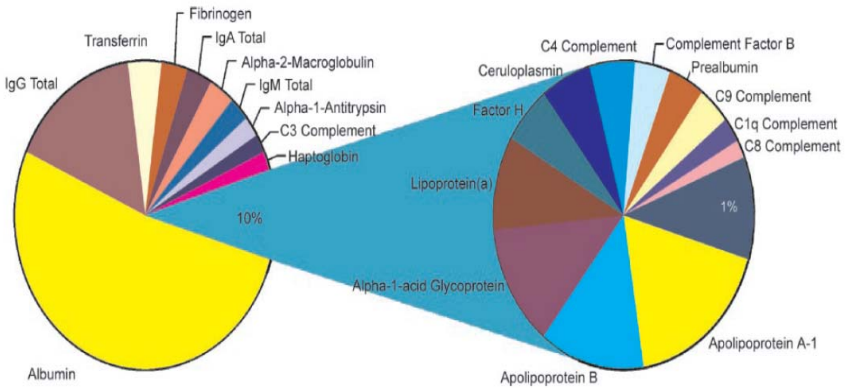


Figure 1.3 **Pie chart representing the relative contribution of proteins within plasma.**

Twenty-two proteins constitute ~99% of the protein content of plasma. [44]

1.3.1.2 Limitations of current proteomic technologies

The three main challenges in proteomic technologies today are poor sensitivity, low resolution, and poor reproducibility. Sensitivity is especially important given the aforementioned dynamic range of proteins. In order to be flagged as a potential biomarker, the low abundance molecular markers must be detected and quantified in the presence of an overwhelming presence of peptides derived from the most abundant proteins. Unfortunately, these low abundance markers are almost always below the limit of detection of current assays designed for unbiased biomarker discovery. Although contemporary mass spectrometers can achieve attomolar sensitivities for the detection of isolated compounds, their working dynamic range typically spans only three orders of magnitude within a single mass spectrum. Significant ion suppression of lower abundance analytes in plasma masks ion signals of less abundant species with similar mass-to-charge (*m/z*) ratios (Fig. 1.4). Low resolution instruments with poor mass accuracy also drastically limit the sensitive and specific detection of low abundance analytes.

Due to the presence of protein isoforms, a separation technique with high resolving power is necessary to simplify the proteome into simpler mixtures for sufficient depth of

coverage of complex samples. As such, most separation techniques are multidimensional, capitalizing on the different physicochemical properties of proteins. As a consequence, a single sample is often separated into tens of fractions, each requiring several hours of on-instrument analysis time, markedly limiting sample throughput.

Reproducibility is a formidable challenge in proteomics as exemplified by the high number of potential protein and peptide biomarkers discovered by scientists from various labs, many of which are non-overlapping for the same disease investigated. Demonstration of reproducibility is crucial to impart confidence on the discovered biomarkers and on the assay employed.

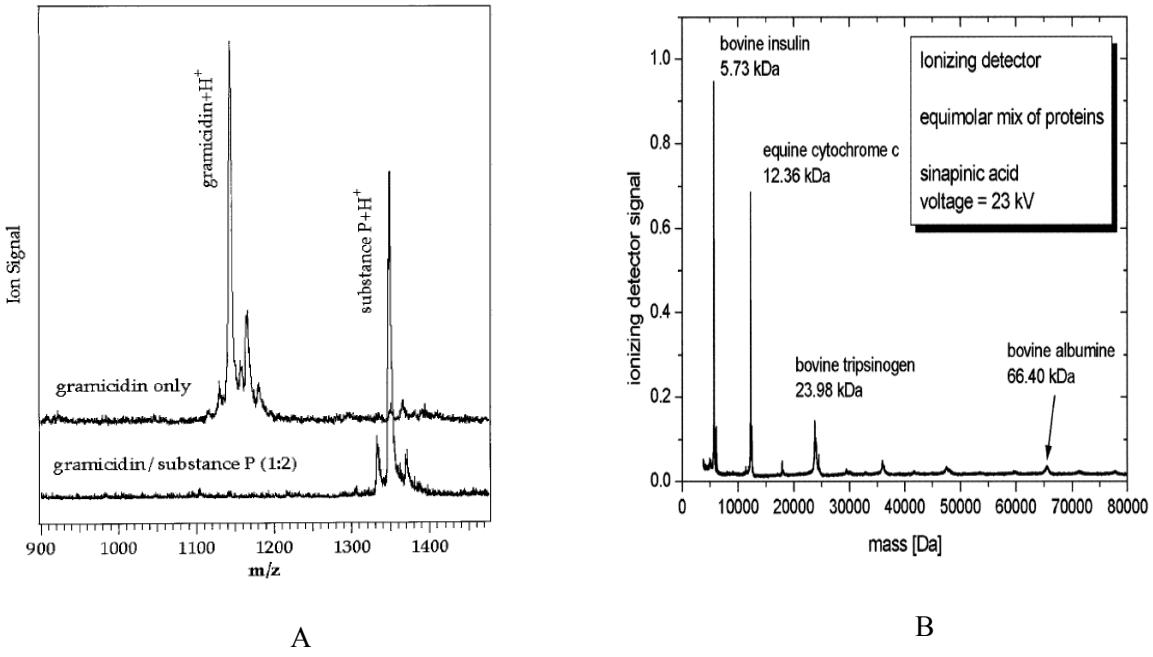


Figure 1.4 **Ion suppression effect in MALDI TOF MS.** (A) Mass spectra of gramicidin S and of a 1:2 mixture of gramicidin S and substance P in dihydroxybenzoic acid matrix. Gramicidin S concentration is the same in both spectra, taken under identical conditions. [45] (B) A mass spectrum of a sample consisting of four protein standards in equimolar concentrations in sinapinic acid matrix. [46]

Unbiased protein biomarker discovery in human plasma has been largely unsuccessful to date. This is reflected in the disappointing performance of attempts at large-scale characterization of the human plasma proteome. For instance, the recent reanalysis of the 3,020 proteins initially identified by HUPO only resulted in 889 proteins identified with at least a 95% confidence level [47, 48]. This suggests that <10% of the core plasma proteome is being effectively sampled with current approaches, a small fraction biased towards proteins of higher abundance (>1 μg/ml). Thus, even with substantial improvements in sensitivity and mass accuracy over the past decade, there is a profound mismatch between complex biological protein mixtures and the capabilities of the MS instrumentation used to analyze them.

1.3.1.3 The curse of heterogeneity

In addition to the technological limitations, the inherent variability in the proteomes of different individuals confounds the unbiased discovery of new biomarkers. It is critical to select extremely well phenotyped individuals in discovery studies and to control for a great variety of factors (e.g. age, ethnicity, time of sample collection) when obtaining samples from a disease and control population. This is inherently difficult.

1.4 PROTEOMICS BIOMARKER DISCOVERY IN AUTOIMMUNE DISEASES

Autoimmune diseases occur in up to 3-5% of the general population [49]. Autoimmunity (Fig. 1.5) is a consequence of the breakdown of the body's self tolerance protection mechanisms where immune molecules launch an attack on self molecules mistakenly deemed to be non-self, culminating in inflammation and tissue damage.

Based on the 'single initiating antigen' hypothesis, all autoimmune diseases are initiated by a response to a single antigen. As the disease progresses, the response broadens through the process of determinant spreading to include other parts of the same molecule and other antigens

in the same tissue, culminating in organ-specific and systemic autoimmune diseases (Table 1.1). Environmental factors that could contribute to autoimmune reactions include pathogenic (bacterial or viral) exposure capable of inducing molecular mimicry of self-antigens for self-reactive T cell recognition, change of physiological state as in pregnancy which affects the hormonal status, or lifestyle activities such as smoking and diet [49].

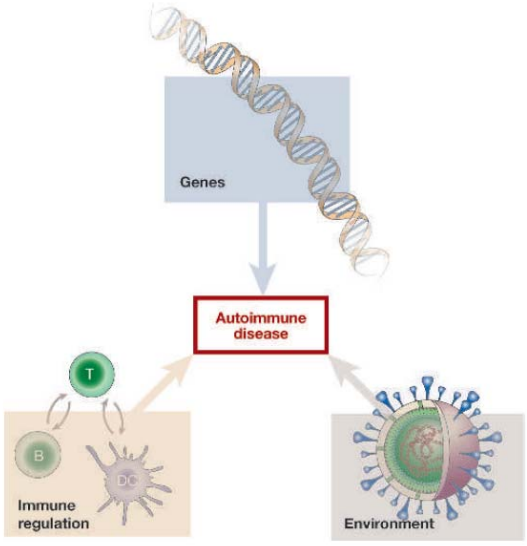


Figure 1.5 **Requirements for the development of autoimmune disease.** The environment can trigger autoimmunity in genetically predisposed individuals under conditions of immune dysregulation. [50]

Disease	Organ	Examples of known autoantigens	Mechanism of damage	Prevalence (%)
Organ-specific autoimmune diseases				
Thyroiditis (autoimmune)	thyroid	thyroglobulin, thyroid peroxidase	T cells/antibody	1.0–2.0
Gastritis	stomach	H^+/K^+ ATPase, intrinsic factor	T cells/antibody	1–2% in > 60-y-old
Celiac disease	small bowel	transglutaminase	T cells/antibody	0.2–1.1
Graves disease	thyroid	thyroid-stimulating hormone receptor	antibody	0.2–1.1
Vitiligo	melanocytes	tyrosinase, tyrosinase-related protein-2	T cells/antibody	0.4
Type 1 diabetes	pancreas β cells	insulin, glutamic acid decarboxylase	T cells	0.2–0.4
Multiple sclerosis	brain/spinal cord	myelin basic protein, proteolipid protein	T cells	0.01–0.15
Pemphigus	skin	desmogleins (for example, desmoglein 1)	antibody	< 0.01 – > 3.0
Hepatitis (autoimmune)	liver	hepatocyte antigens (cytochrome P450)	T cells/antibody	< 0.01
Myasthenia gravis	muscle	acetylcholine receptor	antibody	< 0.01
Primary biliary cirrhosis	liver bile ducts	2-oxoacid dehydrogenase complexes	T cells/antibody	< 0.01
Systemic autoimmune diseases				
Rheumatoid arthritis	joints, lungs, heart etc.	IgG, filaggrin, fibrin etc.	T cells in joint?/antibody	0.8
Systemic lupus (SLE)	skin, joints, kidneys brain, lungs, heart, others	nuclear antigens (DNA, histones, ribonucleoproteins), others	antibody	0.1
Polymyositis/dermato-myositis	skeletal muscle (predominant) lungs, heart, joints, others	muscle antigens, aminoacyl-tRNA synthetases, other nuclear antigens	T cells/antibody	< 0.01

Diseases are listed by category (organ-specific versus systemic) and then by prevalence. Unless referred to in this review, only diseases with a prevalence of greater than 0.1% (> than 1 in 1000) are included in this table. Diseases without a known antigenic target such as inflammatory bowel disease (ulcerative colitis and Crohns disease) or spondyloarthropathies, are also not included. Many of the papers on the prevalence of these diseases have been reviewed[1]. Other sources were also used[86–88].

Table 1.1 Examples of organ-specific and systemic autoimmune diseases with known autoantigen targets. [49]

1.4.1 Correlation to the HLA system

Major histocompatibility complex (MHC) Class II molecules have long been implicated as contributors to the genetic basis of autoimmunity [51]. The human MHC gene region encodes multiple HLA molecules and a number of other immunologically active molecules (Fig. 1.6). The alleles encoding MHC Class II proteins control the antigen specificity of the autoimmune response and therefore are critical determinants of immune activation. Different alleles might have different abilities to present peptide from the target cells to autoreactive $CD4^+$ T cells. Certain Class II alleles might even predispose to positive selection and reduce negative selection of self-reactive T cells in the thymus during T cell maturation.

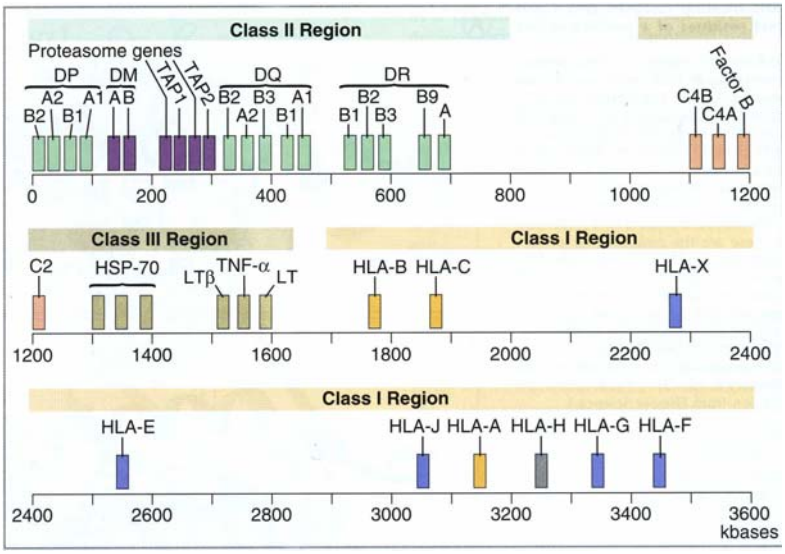


Figure 1.6 **Map of the human MHC**. C4, C2, B, complement proteins; DM, TAP, proteasomes, protein involved in antigen processing; HLA, human leukocyte antigen; HSP, heat shock protein; LT, TNF, cytokines. [52]

A large number of susceptibility genes are usually implicated in common autoimmune diseases. Associations between particular MHC Class II alleles and autoimmune diseases are well documented [53, 54], especially through linkage disequilibrium, a phenomenon where some of the alleles expressed at one Class-II locus are found linked frequently to specific alleles at a neighboring locus. As a result, this confounds the association between a disease and a particular locus as it may reflect an effect of either the locus studied, an adjacent genetic locus, or a combination of both. This renders disease gene identification extremely challenging. The utility of genetic markers in disease prediction is limited by its low specificity where many genetically predisposed individuals do not develop the disease. HLA-associated autoimmune diseases are complex, arising as a result of the interaction of environmental influences with a polygenic background of susceptibility. These environmental triggers result in the alteration of the protein

constituents of the system, opening up the opportunity for autoimmune disease biomarker discovery via proteomics.

1.4.2 Current proteomics studies in autoimmune diseases

It has so far proven extremely difficult to develop novel protein biomarkers for autoimmune diseases. Due to the heterogeneity in clinical presentation and disease course, new multiparameter assays with improved sensitivity and specificity over current single biomarkers are gravely needed to detect the onset of autoimmune diseases at an early stage.

Proteomic autoimmune studies have been reported on diseases such as Type I diabetes [55, 56], collagen-induced arthritis [57], rheumatoid arthritis (RA) [58-60], celiac disease [61], multiple sclerosis [62, 63], and lupus [64]. As is true of proteomic studies in general, the low throughput 2DGE-MS approach (described below) has been the method of choice.

1.5 PROTEOMICS BIOMARKER DISCOVERY MODALITIES

To date, the two major unbiased technology platform adopted in proteomic studies are two-dimensional gel electrophoresis (2DGE) and MS. The dynamic range of 2DGE is 10^4 whereas that of MALDI MS is about 10^3. This is still 6 to 7 magnitudes less than necessary to fully probe the blood proteome. In an attempt to increase resolution, most analytical configurations involve combination of existing technologies, since none of the existing separation and identification methodologies alone can provide a full account of the protein composition in a complex mixture. Each has its own strengths and limitations.

1.5.1 Two-dimensional gel electrophoresis (2DGE)

A popular approach in proteomic profiling experiments is the 2DGE-MS configuration (Fig. 1.7). Here, isoelectric focusing based on charges in the first dimension and mass separation in the second dimension are performed. Subsequently, protein spots with differential staining are excised, digested, and analyzed by MS [65-67]. Its advantages include information on PTMs and protein processing from its size and charged forms. Albeit successful to a certain extent, this technique suffers from low staining sensitivity and low throughput. It is time-consuming to pick all spots of interest and difficult to perform automatic data analysis to acquire the large number of 2D gel images. Resolution is limited by protein heterogeneity that leads to spot overlap and obscuring of low abundance proteins that comigrate with the high abundance ones. A spot could represent one or a few proteins with specific charges and similar mass. The limited loading capacity means that 2DGE is biased to high abundance proteins. Smaller proteins, extremely acidic, basic, or hydrophobic proteins such as transmembrane proteins are most often underrepresented. Reproducibility is also a concern due to the variability between gel runs that confound spot matching across gels. The average coefficient of variation (CV) for 2DGE runs is around 20% for plasma samples and around 6% for cerebrospinal fluid (CSF) samples.

A variation of this approach is called two-dimensional difference gel electrophoresis (2D-DIGE) [68]. Here, the two sample groups to be compared are labeled separately with distinct fluorescent dyes, combined and run in a single gel. The dye-labeled samples are then viewed individually by scanning the gel at different wavelengths to obtain quantitative information. Although it circumvents the problem of spot matching between gels, it still suffers from low resolution and low throughput.

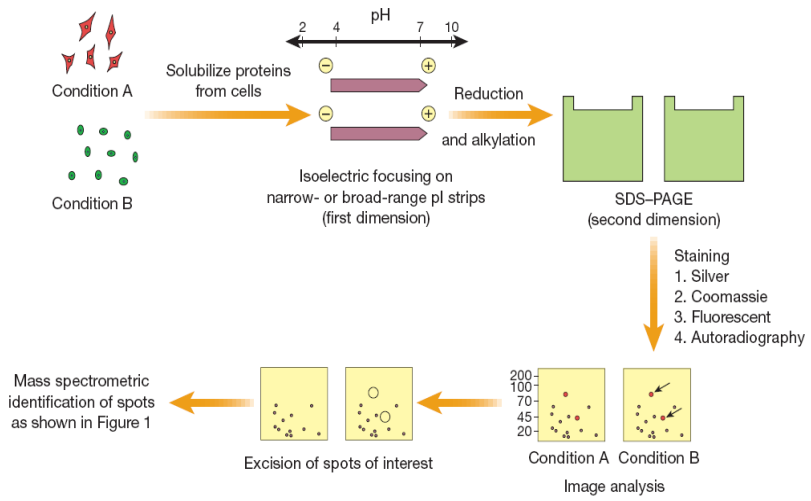


Figure 1.7 **A schematic showing the two-dimensional gel approach**. Two different samples, A and B, are applied to a 'first dimension' gel strip that separates the proteins based on their isoelectric points. Then, the strip is applied to a 'second dimension' SDS–PAGE gel where proteins are denatured and separated on the basis of size. After staining, the resulting protein spots are quantified. Differential spots are then excised and subject to MS analysis. [69]

1.5.2 Multidimensional protein identification technology (MudPIT)

The favored unbiased, MS-based approach is MudPIT (Fig. 1.8) which involves the sequential, multidimensional column separation of the enzymatically digested proteome of interest by strong cation exchange and reverse phase before on-line introduction of the eluted fractions into the mass spectrometer for analysis [70]. This is followed by algorithmic identification of the protein fragments based on the mass spectral data. Protein identification is achieved through peptide mass fingerprinting based on the correlation between the experimentally observed peptide masses and the theoretical spectra of *in silico* digests of proteins listed in databases using specialized softwares such as MASCOT or SEQUEST. Alternatively, proteins can be identified via tandem MS (MS/MS) peptide sequencing. Peptide sequencing is based on induction of random cleavage

of peptide bonds between adjacent amino acids by collision with an inert gas (e.g. nitrogen, helium, or argon) of the parent ion to produce fragment ions. Database searches can be performed using the product-ion mass spectra alone or in concert with the molecular weight of the parent ion to increase confidence.

This powerful technique has proven successful in increasing the coverage of the yeast and blood proteomes [71-75]. The drawbacks of this approach are the risk of cross-contamination between samples and its high cost in throughput. It also requires a large amount of protein to begin with, which precludes its routine use with specimens such as scarce clinical samples. The amount of sample that can be loaded becomes limited when capillary columns are used for better sensitivity in liquid chromatography-tandem MS (LC-MS/MS).

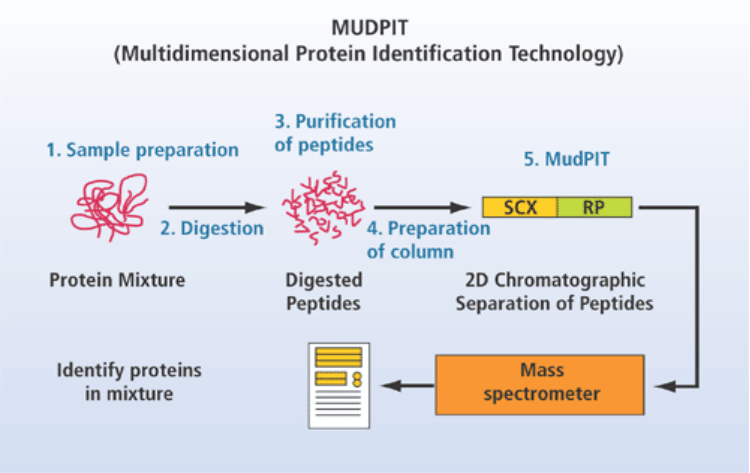


Figure 1.8 **Workflow for multidimensional protein identification technology (MudPIT)**. Proteins are prepared, digested into constituent peptides, which are then separated by 2D chromatography and analyzed via tandem mass spectrometry. SCX = strong cationic exchange; RP = reversed-phase.

1.5.3 Novel high-throughput biomarker discovery modalities

The initial discovery phase of biomarker development has to be conducted with a statistically significant number of representative samples of the disease of interest while the rigorous validation phase calls for an even larger number of samples from the population incorporating a broader range of cases and controls to capture the environmental, genetic, biological and stochastic variation in the population to be tested. The 2DGE and MudPIT technologies mentioned above are not, by nature, designed to handle large number of samples. Therefore, they are more suitable for initial discovery efforts than for larger clinical validations. A new high-throughput methodology that allows a facile transition from discovery to validation will significantly reduce the translational timeline from marker development to introduction in the clinical setting. The methodology will also have to find a compromise between the need for depth and comprehensiveness of sample proteome analysis, and the need for large sample numbers.

Although still at their infancy, novel high-throughput protein profiling technologies are slowly making their way into the mainstream of proteomics studies. An up and coming technology utilizes the protein microarray platform to probe for autoantigens and/or autoantibodies in autoimmune samples. A hallmark of many autoimmune diseases is the presence of high-affinity, high-avidity autoantibodies. Antibodies have long been used for the diagnosis and classification of autoimmune disease [76]. Although still in its early stages, impressive studies using protein microarrays have shown great utility in discovering new autoantigens and in differential pattern profiling [77-81]. This platform is beyond the scope of this dissertation and will not be discussed further.

Another notable high-throughput approach is the SELDI TOF MS technology which has garnered much interest since its introduction in the 1990s [82]. This platform is discussed below and forms the basis of the methodology described in Chapter 2.

1.6 SELDI TOF MS

A mass spectrometer is comprised of three basic components: an ion source that converts analytes into gaseous phase ions, a mass analyzer that separates the ionized species based on their mass-to-charge (*m/z*) ratios, and a detector that registers the number of ions at each *m/z* value. All mass spectrometers, regardless of ionization mode or mass analyzer used, output mass spectra which plot the signal intensity of the ions produced against their *m/z* ratios. The ionization process of an analyte is dependent on its intrinsic physicochemical properties which govern its ionization efficiency, a measure of how likely the analyte will be ionized and detected by the mass spectrometer. In the presence of molecules that have a higher affinity for proton sequestration from the ionizing matrix, the probability of the outcompeted molecules to be represented in the mass spectrum is lowered significantly. This phenomenon, known as ion suppression, can be advantageous in simplifying the complex proteome to be interrogated to just the 'ionizable subproteome'.

A protein profiling approach that relies solely on the comparison of the signal intensity of these mass peaks from both disease and non-disease samples to uncover candidate biomarkers has been making its way into the mainstream of proteomics studies. These peaks, regardless of whether they are identified or not, will qualify for disease marker candidacy as long as the variation between the two states compared is consistent and reproducible. MALDI TOF MS has been the preferred ionization technique for differential protein pattern profiling studies as it can readily be automated to handle large number of samples in a short time frame. However, due to its limited dynamic range and the effect of ion suppression, analysis of complex samples still suffer from limited depth of coverage and low representation of the low abundance proteins in the absence of some form of pre-fractionation. A notable variation of this approach is surface-enhanced laser desorption-ionization (SELDI) TOF MS, where on-chip pre-fractionation is coupled to MALDI TOF MS.

1.6.1 Fundamentals

SELDI TOF MS combines surface retentate chromatographic separation of proteins on a ProteinChip array to the direct analysis by mass spectrometry. The surface of the array chip is coated with capture agents of varying chromatographic properties, including anion exchange, cation exchange, normal and reverse phase, and metal affinity (Fig. 1.9).

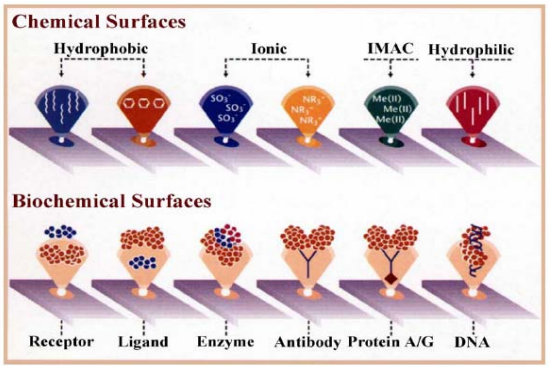


Figure 1.9 Surface chemistries available on ProteinChip arrays.

Complex biological samples, such as serum or CSF, are incubated on the chip until binding equilibrium is reached. Only a subset of the proteins (depending on the surface chemistry employed) in the sample binds to the chromatographic surface of the chip, and the unbound or weakly bound proteins are washed away. The bait region (spots on the chip with immobilized capture agents) containing individual captured protein samples is then overlaid with a coating of organic acid matrix (e.g. α-cyano-4-hydroxycinnamic acid, CHCA) which co-crystallizes and embeds the proteins. Then, the entire chip is introduced into the vacuum chamber of a mass spectrometer and each spot is ablated with a focused laser beam (at 337nm for nitrogen laser, 355nm for Nd:YAG). The excess organic matrix which serves as an energy transfer medium undergoes instant sublimation, liberating the embedded protein molecules in the process, to form ions in the gas phase. After numerous ion-molecule collisions in the plume, single protonated

protein ions are formed. The positively charged ions are propelled forward in vacuum into a flight tube by the high positive voltage applied to the chip. The mass-to-charge ratio of each ion is estimated from the time it takes for the launched ion to reach the detector electrode at the end of the flight tube. The time it takes the ion to reach the detector is dictated by its mass, with smaller ions traveling faster. Consequently, the spectrum provides a 'time-of-flight' (TOF) signature of ions ordered by size (Ion signal intensity versus *m/z*) (Fig. 1.10).

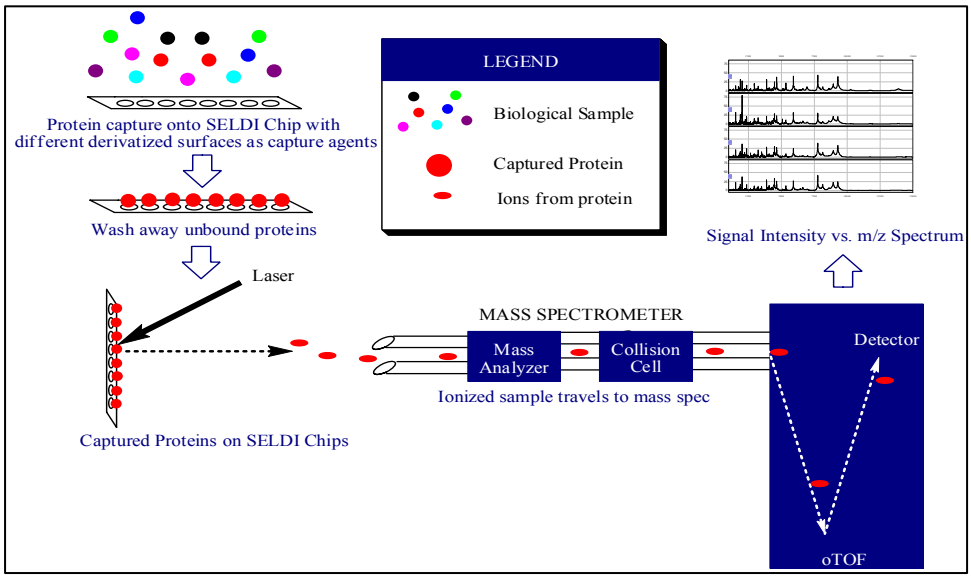


Figure 1.10 **Schematic showing the operation of SELDI TOF MS**. Orthogonal TOF used for detection confers greater resolution of protein/peptide species.

Protein chips offer the advantage of proteome simplification by coupling retentate chromatography to mass spectrometry. It enables the simultaneous analysis of thousands of proteins while consuming minute amount of samples compared with the more traditional 2DGE, facilitating the identification both of individual biomarkers and diagnostic protein patterns. It allows for high-throughput, rapid analysis of samples (minutes or hours for large sample sets as opposed to days for 2DGE and MudPIT analyses), generation of simple mass spectrum for

analysis since ions are predominantly singly charged, high detection sensitivity down to a few attomoles for short peptides and obviates the need for labeling molecules.

In comparison to the MudPIT shotgun approach where analysis is performed at the peptide level, SELDI TOF MS enables protein-level analysis. Information such as the integrity of the original protein, its isoforms, and PTMs are preserved. MALDI TOF MS, from which SELDI is derived, is biased towards peptides and proteins predominantly in the low molecular weight (LMW) mass range. This is, in part, due to diminishing ionization efficiency with an increase in mass. Therefore, for the best possible reproducibility and mass accuracy using SELDI, high-resolution MS is mandatory. Proteomic diagnostic pattern "fingerprint" analysis with SELDI begins with high dimensional data, where thousands of proteins are represented by peaks on the spectra and within this forest of ion peaks, identify patterns of LMW proteins/peptides as the diagnostic itself. Patterns emerging from this multiparametric analysis of mass peaks can potentially be of high specificity and amenable for immediate validation on blinded statistically significant study sets.

1.6.2 SELDI studies

Since its conception, SELDI has been adopted predominantly by the cancer research community [83-90]. The past few years have witnessed the diversification of its applications to include fields such as immunology [91, 92], neurology [93, 94], and toxicology [95], and across a myriad of biological samples such as plasma, serum, CSF, urine, gastric juices and saliva [96-99]. Although not mainstream, there is also precedence of this platform with promising outcomes in autoimmune and neurodegeneration studies such as Sjögren Syndrome [100, 101], RA [102, 103], and Alzheimer's [94, 104].

1.6.3 SELDI issues

Published reports of the pattern profiles generated by SELDI TOF MS have suggested that this method yields better diagnostic sensitivities and specificities than biomarkers in current use, culminating in extensive publicity [39, 84]. In fact, the results were so groundbreaking that it prompted the US Congress to pass a resolution to urge further funding of the research while the government licensed rights to develop the platform into a commercial diagnostic test for early detection of ovarian cancer [105]. However, initial hype over this technology has been dampened by other reports identifying potential pitfalls [106-111], which, if left unchecked, could lead to false positive patients undergoing unnecessary surgery and false negatives forgoing further screening. These issues encompass all steps from sample procurement through data acquisition to data mining and must be addressed before clinical application can be instituted.

1.6.3.1 Preanalytical variations

The introduction of variability begins at sample collection. Once the type of sample has been determined, care must be taken during the patient selection process. It is imperative that the samples be culled from larger, clinically relevant cohorts to reduce false discovery from biased specimen and be sufficiently large to increase statistical power and avoid unwarranted generalizability [112]. Confounding factors that are unrelated to the disease under study, such as gender, age, and physiological states (e.g. fasting, weight gain/loss, hormonal changes due to pregnancy), must be controlled. Studies of diseases where genetic predisposition is implicated must ensure that the genetic makeup of the patients is well documented.

Unlike DNA, proteins are extremely sensitive to storage, handling, and processing conditions [113, 114]. Minor deviations in sample procurement protocol, such as collection in the fasting or feeding state, or the posture of the patient (e.g. supine or seated position), could affect the analyte concentration by up to 15%, resulting in biased clustering of samples unrelated to the

disease profiled [115]. Many analytes, such as IL-6, exhibit very significant circadian rhythm and, therefore, samples for their measurement must be collected at a set time. Short-and long-term storage conditions are also of paramount importance as the utility of the protein will likely be assessed from stored clinical samples. As an example, CRP is a very stable protein (up to 20 years at –20°C) while tumor necrosis factor is rather labile, requiring sample collection on ice and storage at –70°C or in liquid nitrogen. Other potential factors include the type of sample tubes used, coagulation time for serum, and number of freeze/thaw cycles [116-119]. Standardization of specimen collection and handling has been initiated to minimize bias from these preanalytical variables [120].

1.6.3.2 Analytical variations

The main criticism of SELDI is in its low reproducibility. Both biological and analytical variability affect the reliability of the measurement for diagnosis. The robustness of the methodology has to be demonstrated through reproducibility of protein patterns during data acquisition across different batches of chips, different operators, different sites, and different instrumentation. This is highly dependent on the performance of the mass spectrometer used. Low resolution mass spectrometers that do not compensate for the broad energy spread during the desorption process in MALDI can result in broad peaks with shoulders that are difficult to reproduce, confounding downstream data analysis. Chemical noise from interfering matrix peaks is also a source of dissonance in mass spectra. Low mass accuracy across spectra will generate irreproducible profiles. Poor analytical sensitivity is also a concern particularly when the analyte is present in trace amounts in complex biological materials containing high-abundance molecules. Pre-fractionation is an absolute requirement to uncover diagnostic markers in the ng/ml concentration given the dynamic range of current mass spectrometers. The ideal separation technology that minimizes ion suppression and optimizes the display of mass peaks and, by extension, the number of proteins represented in the spectra will increase the chances of

discovering potential markers. In addition, it also has to be compatible with MS and not involve extensive preparative steps so as to minimize operator bias. The analytical performance of SELDI must be improved such that sensitivity, specificity, and dynamic range can approximate those of diagnosis platforms currently in use if it is to be of clinical utility.

1.6.3.3 Postanalytical variations

The diagnostic potency of the protein profiles is also affected by bioinformatics artifacts during data mining. Data overfitting is a great concern as differential mass peaks that are biased to a particular algorithm during disease model building do not possess true diagnostic prowess but instead are incorporated by chance. As a result, a high false positive rate will ensue when they are applied to actual test sets. Moreover, when these peaks are pursued further for identification purposes, not only will they drain scarce resources and time but the futile efforts will lead no closer to biomarkers that can be developed into a diagnostic test. These statistical challenges inherent in high-dimensionality data sets populated by comparatively few samples are here to stay in unbiased protein profiling studies, an unfortunate inheritance from transcription profiling studies. A more robust data analysis approach has to be adopted to overcome this impedance and reestablish confidence in the candidate marker peaks.

1.7 BIBLIOGRAPHY

1. Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA *et al*: **The sequence of the human genome**. *Science (New York, NY* 2001, **291**(5507):1304-1351.
2. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W *et al*: **Initial sequencing and analysis of the human genome**. *Nature* 2001, **409**(6822):860-921.
3. Clamp M, Fry B, Kamal M, Xie X, Cuff J, Lin MF, Kellis M, Lindblad-Toh K, Lander ES: **Distinguishing protein-coding and noncoding genes in the human genome**. *Proceedings of the National Academy of Sciences of the United States of America* 2007, **104**(49):19428-19433.
4. Birney E, Stamatoyannopoulos JA, Dutta A, Guigo R, Gingeras TR, Margulies EH, Weng Z, Snyder M, Dermitzakis ET, Thurman RE *et al*: **Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project**. *Nature* 2007, **447**(7146):799-816.
5. Wilkins MR, Pasquali C, Appel RD, Ou K, Golaz O, Sanchez JC, Yan JX, Gooley AA, Hughes G, Humphery-Smith I *et al*: **From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis**. *Bio/technology (Nature Publishing Company)* 1996, **14**(1):61-65.
6. Huber LA: **Is proteomics heading in the wrong direction?** *Nat Rev Mol Cell Biol* 2003, **4**(1):74-80.
7. Fields S: **Proteomics. Proteomics in genomeland**. *Science (New York, NY* 2001, **291**(5507):1221-1224.

8. Tanaka K, Waki H, Ido Y, Akita S, Yoshida Y, Yoshida T: **Protein and polymer analyses up to m/z 100 000 by laser ionization time-of-flight mass spectrometry**. *Rapid communications in mass spectrometry* 1988, **2**(8):151-153.

9. Karas M, Hillenkamp F: **Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons**. *Analytical chemistry* 1988, **60**(20):2299-2301.

10. Fenn JB, Mann M, Meng CK, Wong SF, Whitehouse CM: **Electrospray ionization for mass spectrometry of large biomolecules**. *Science (New York, NY* 1989, **246**(4926):64-71.

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

12. Griffin TJ, Gygi SP, Ideker T, Rist B, Eng J, Hood L, Aebersold R: **Complementary profiling of gene expression at the transcriptome and proteome levels in Saccharomyces cerevisiae**. *Mol Cell Proteomics* 2002, **1**(4):323-333.

13. Gygi SP, Rochon Y, Franza BR, Aebersold R: **Correlation between protein and mRNA abundance in yeast**. *Molecular and cellular biology* 1999, **19**(3):1720-1730.

14. Lee PS, Shaw LB, Choe LH, Mehra A, Hatzimanikatis V, Lee KH: **Insights into the relation between mrna and protein expression patterns: II. Experimental observations in Escherichia coli**. *Biotechnology and bioengineering* 2003, **84**(7):834-841.

15. Parekh RB, Rohlff C: **Post-translational modification of proteins and the discovery of new medicine**. *Current opinion in biotechnology* 1997, **8**(6):718-723.

16. Lu P, Vogel C, Wang R, Yao X, Marcotte EM: **Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation**. *Nature biotechnology* 2007, **25**(1):117-124.

17. Longstreth WT, Jr., Koepsell TD, Ton TG, Hendrickson AF, van Belle G: **The epidemiology of narcolepsy**. *Sleep* 2007, **30**(1):13-26.

18. Lee JW, Devanarayan V, Barrett YC, Weiner R, Allinson J, Fountain S, Keller S, Weinryb I, Green M, Duan L *et al*: **Fit-for-purpose method development and validation for successful biomarker measurement**. *Pharmaceutical research* 2006, **23**(2):312-328.

19. Ramaswamy S, Perou CM: **DNA microarrays in breast cancer: the promise of personalised medicine**. *Lancet* 2003, **361**(9369):1576-1577.

20. Fernie AR, Trethewey RN, Krotzky AJ, Willmitzer L: **Metabolite profiling: from diagnostics to systems biology**. *Nat Rev Mol Cell Biol* 2004, **5**(9):763-769.

21. Menon U, Jacobs IJ: **Recent developments in ovarian cancer screening**. *Current opinion in obstetrics & gynecology* 2000, **12**(1):39-42.

22. Atkinson AJ, Colburn WA, DeGruttola VG, DeMets DL, Downing GJ, Hoth DF, Oates JA, Peck CC, Schooley RT, Spilker BA *et al*: **Biomarkers and surrogate endpoints: preferred definitions and conceptual framework**. *Clinical pharmacology and therapeutics* 2001, **69**(3):89-95.

23. Anderson NL, Anderson NG: **The human plasma proteome: history, character, and diagnostic prospects**. *Mol Cell Proteomics* 2002, **1**(11):845-867.

24. Gutman S, Kessler LG: **The US Food and Drug Administration perspective on cancer biomarker development**. *Nature reviews* 2006, **6**(7):565-571.

25. Rifai N, Gillette MA, Carr SA: **Protein biomarker discovery and validation: the long and uncertain path to clinical utility**. *Nature biotechnology* 2006, **24**(8):971-983.

26. McLerran D, Grizzle WE, Feng Z, Thompson IM, Bigbee WL, Cazares LH, Chan DW, Dahlgren J, Diaz J, Kagan J *et al*: **SELDI-TOF MS whole serum proteomic profiling with IMAC surface does not reliably detect prostate cancer**. *Clinical chemistry* 2008, **54**(1):53-60.

27. Rosen DG, Wang L, Atkinson JN, Yu Y, Lu KH, Diamandis EP, Hellstrom I, Mok SC, Liu J, Bast RC, Jr.: **Potential markers that complement expression of CA125 in epithelial ovarian cancer**. *Gynecologic oncology* 2005, **99**(2):267-277.

28. Thompson IM, Pauler DK, Goodman PJ, Tangen CM, Lucia MS, Parnes HL, Minasian LM, Ford LG, Lippman SM, Crawford ED *et al*: **Prevalence of prostate cancer among men with a prostate-specific antigen level < or =4.0 ng per milliliter**. *The New England journal of medicine* 2004, **350**(22):2239-2246.

29. Beduschi MC, Oesterling JE: **Percent free prostate-specific antigen: the next frontier in prostate-specific antigen testing**. *Urology* 1998, **51**(5A Suppl):98-109.

30. Mor G, Visintin I, Lai Y, Zhao H, Schwartz P, Rutherford T, Yue L, Bray-Ward P, Ward DC: **Serum protein markers for early detection of ovarian cancer**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(21):7677-7682.

31. Sabatine MS, Morrow DA, de Lemos JA, Gibson CM, Murphy SA, Rifai N, McCabe C, Antman EM, Cannon CP, Braunwald E: **Multimarker approach to risk stratification in non-ST elevation acute coronary syndromes: simultaneous assessment of troponin I, C-reactive protein, and B-type natriuretic peptide**. *Circulation* 2002, **105**(15):1760-1763.

32. Ross PL, Huang YN, Marchese JN, Williamson B, Parker K, Hattan S, Khainovski N, Pillai S, Dey S, Daniels S *et al*: **Multiplexed protein quantitation in Saccharomyces cerevisiae using amine-reactive isobaric tagging reagents**. *Mol Cell Proteomics* 2004, **3**(12):1154-1169.

33. Ong SE, Blagoev B, Kratchmarova I, Kristensen DB, Steen H, Pandey A, Mann M: **Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics**. *Mol Cell Proteomics* 2002, **1**(5):376-386.

34. Gygi SP, Rist B, Gerber SA, Turecek F, Gelb MH, Aebersold R: **Quantitative analysis of complex protein mixtures using isotope-coded affinity tags**. *Nature biotechnology* 1999, **17**(10):994-999.

35. Kingsmore SF: **Multiplexed protein measurement: technologies and applications of protein and antibody arrays**. *Nat Rev Drug Discov* 2006, **5**(4):310-320.

36. de Wildt RM, Mundy CR, Gorick BD, Tomlinson IM: **Antibody arrays for high-throughput screening of antibody-antigen interactions**. *Nature biotechnology* 2000, **18**(9):989-994.

37. Sihlbom C, Kanmert I, Bahr H, Davidsson P: **Evaluation of the combination of bead technology with SELDI-TOF-MS and 2-D DIGE for detection of plasma proteins**. *Journal of proteome research* 2008, **7**(9):4191-4198.

38. Lowenthal MS, Mehta AI, Frogale K, Bandle RW, Araujo RP, Hood BL, Veenstra TD, Conrads TP, Goldsmith P, Fishman D *et al*: **Analysis of albumin-associated peptides and proteins from ovarian cancer patients**. *Clinical chemistry* 2005, **51**(10):1933-1945.

39. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, Steinberg SM, Kohn EC, Fishman DA, Whitely G *et al*: **High-resolution serum proteomic features for ovarian cancer detection**. *Endocrine-related cancer* 2004, **11**(2):163-178.

40. Han MH, Hwang SI, Roy DB, Lundgren DH, Price JV, Ousman SS, Fernald GH, Gerlitz B, Robinson WH, Baranzini SE *et al*: **Proteomic analysis of active multiple sclerosis lesions reveals therapeutic targets**. *Nature* 2008, **451**(7182):1076-1081.

41. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA: **Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution**. *Mol Cell Proteomics* 2007, **6**(12):2212-2229.

42. Burtis CA, Ashwood ER, Bruns DE: **Tietz Textbook of Clinical Chemistry and Molecular Diagnostics**, 4 edn. Philadelphia: Elsevier Science; 2005.

43. Sturgeon CM, Hoffman BR, Chan DW, Ch'ng SL, Hammond E, Hayes DF, Liotta LA, Petricoin EF, Schmitt M, Semmes OJ *et al*: **National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for use of tumor markers in clinical practice: quality requirements**. *Clinical chemistry* 2008, **54**(8):e1-e10.

44. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD: **Characterization of the low molecular weight human serum proteome**. *Mol Cell Proteomics* 2003, **2**(10):1096-1103.

45. Knochenmuss R, Stortelder A, Breuker K, Zenobi R: **Secondary ion-molecule reactions in matrix-assisted laser desorption/ionization**. *J Mass Spectrom* 2000, **35**(11):1237-1245.

46. Twerenbold D, Gerber D, Gritti D, Gonin Y, Netuschill A, Rossel F, Schenker D, Vuilleumier JL: **Single molecule detector for mass spectrometry with mass independent detection efficiency**. *Proteomics* 2001, **1**(1):66-69.

47. States DJ, Omenn GS, Blackwell TW, Fermin D, Eng J, Speicher DW, Hanash SM: **Challenges in deriving high-confidence protein identifications from data gathered by a HUPO plasma proteome collaborative study**. *Nature biotechnology* 2006, **24**(3):333-338.

48. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS *et al*: **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 3020 proteins and a publicly-available database**. *Proteomics* 2005, **5**(13):3226-3245.

49. Marrack P, Kappler J, Kotzin BL: **Autoimmune disease: why and where it occurs**. *Nat Med* 2001, **7**(8):899-905.

50. Fathman CG, Soares L, Chan SM, Utz PJ: **An array of possibilities for the study of autoimmunity**. *Nature* 2005, **435**(7042):605-611.

51. Nepom GT, Erlich H: **MHC class-II molecules and autoimmunity**. *Annual review of immunology* 1991, **9**:493-525.

52. Abbas AK, Lichtman AH: **Cellular and Molecular Immunology**, Fifth edn: Saunders; 2005.

53. Nepom GT, Kwok WW: **Molecular basis for HLA-DQ associations with IDDM**. *Diabetes* 1998, **47**(8):1177-1184.

54. Nepom GT: **Major histocompatibility complex-directed susceptibility to rheumatoid arthritis**. *Advances in immunology* 1998, **68**:315-332.

55. Sparre T, Christensen UB, Mose Larsen P, Fey SJ, Wrzesinski K, Roepstorff P, Mandrup-Poulsen T, Pociot F, Karlsen AE, Nerup J: **IL-1beta induced protein changes in diabetes prone BB rat islets of Langerhans identified by proteome analysis**. *Diabetologia* 2002, **45**(11):1550-1561.

56. Larsen PM, Fey SJ, Larsen MR, Nawrocki A, Andersen HU, Kahler H, Heilmann C, Voss MC, Roepstorff P, Pociot F *et al*: **Proteome analysis of interleukin-1beta--induced changes in protein expression in rat islets of Langerhans**. *Diabetes* 2001, **50**(5):1056-1063.

57. Lorenz P, Bantscheff M, Ibrahim SM, Thiesen HJ, Glocker MO: **Proteome analysis of diseased joints from mice suffering from collagen-induced arthritis**. *Clin Chem Lab Med* 2003, **41**(12):1622-1632.

58. Hueber W, Tomooka BH, Zhao X, Kidd BA, Drijfhout JW, Fries JF, van Venrooij WJ, Metzger AL, Genovese MC, Robinson WH: **Proteomic analysis of secreted proteins in early rheumatoid arthritis: anti-citrulline autoreactivity is associated with up regulation of proinflammatory cytokines**. *Ann Rheum Dis* 2007, **66**(6):712-719.

59. Sinz A, Bantscheff M, Mikkat S, Ringel B, Drynda S, Kekow J, Thiesen HJ, Glocker MO: **Mass spectrometric proteome analyses of synovial fluids and plasmas from**

patients suffering from rheumatoid arthritis and comparison to reactive arthritis or osteoarthritis. *Electrophoresis* 2002, **23**(19):3445-3456.

60. Drynda S, Ringel B, Kekow M, Kuhne C, Drynda A, Glocker MO, Thiesen HJ, Kekow J: **Proteome analysis reveals disease-associated marker proteins to differentiate RA patients from other inflammatory joint diseases with the potential to monitor anti-TNFalpha therapy**. *Pathol Res Pract* 2004, **200**(2):165-171.

61. Stulik J, Hernychova L, Porkertova S, Pozler O, Tuckova L, Sanchez D, Bures J: **Identification of new celiac disease autoantigens using proteomic analysis**. *Proteomics* 2003, **3**(6):951-956.

62. Almeras L, Lefranc D, Drobecq H, de Seze J, Dubucquoi S, Vermersch P, Prin L: **New antigenic candidates in multiple sclerosis: identification by serological proteome analysis**. *Proteomics* 2004, **4**(7):2184-2194.

63. Dumont D, Noben JP, Raus J, Stinissen P, Robben J: **Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients**. *Proteomics* 2004, **4**(7):2117-2124.

64. Thebault S, Gilbert D, Hubert M, Drouot L, Machour N, Lange C, Charlionet R, Tron F: **Orderly pattern of development of the autoantibody response in (New Zealand White x BXSB)F1 lupus mice: characterization of target antigens and antigen spreading by two-dimensional gel electrophoresis and mass spectrometry**. *J Immunol* 2002, **169**(7):4046-4053.

65. Celis JE, Celis P, Ostergaard M, Basse B, Lauridsen JB, Ratz G, Rasmussen HH, Orntoft TF, Hein B, Wolf H *et al*: **Proteomics and immunohistochemistry define some of the steps involved in the squamous differentiation of the bladder transitional epithelium: a novel strategy for identifying metaplastic lesions**. *Cancer Res* 1999, **59**(12):3003-3009.

66. Celis JE, Wolf H, Ostergaard M: **Bladder squamous cell carcinoma biomarkers derived from proteomics**. *Electrophoresis* 2000, **21**(11):2115-2121.

67. Seliger B, Kellner R: **Design of proteome-based studies in combination with serology for the identification of biomarkers and novel targets**. *Proteomics* 2002, **2**(12):1641-1651.

68. Lilley KS: **Protein profiling using two-dimensional difference gel electrophoresis (2-D DIGE)**. *Current protocols in protein science / editorial board, John E Coligan [et al* 2003, **Chapter 22**:Unit 22 22.

69. Pandey A, Mann M: **Proteomics to study genes and genomes**. *Nature* 2000, **405**(6788):837-846.

70. Link AJ, Eng J, Schieltz DM, Carmack E, Mize GJ, Morris DR, Garvik BM, Yates JR, 3rd: **Direct analysis of protein complexes using mass spectrometry**. *Nat Biotechnol* 1999, **17**(7):676-682.

71. Washburn MP, Wolters D, Yates JR, 3rd: **Large-scale analysis of the yeast proteome by multidimensional protein identification technology**. *Nat Biotechnol* 2001, **19**(3):242-247.

72. Chen EI, Hewel J, Felding-Habermann B, Yates JR, 3rd: **Large scale protein profiling by combination of protein fractionation and multidimensional protein identification technology (MudPIT)**. *Mol Cell Proteomics* 2006, **5**(1):53-56.

73. Deshaies RJ, Seol JH, McDonald WH, Cope G, Lyapina S, Shevchenko A, Verma R, Yates JR, 3rd: **Charting the protein complexome in yeast by mass spectrometry**. *Mol Cell Proteomics* 2002, **1**(1):3-10.

74. Adkins JN, Varnum SM, Auberry KJ, Moore RJ, Angell NH, Smith RD, Springer DL, Pounds JG: **Toward a human blood serum proteome: analysis by multidimensional separation coupled with mass spectrometry**. *Mol Cell Proteomics* 2002, **1**(12):947-955.

75. Shen Y, Jacobs JM, Camp DG, 2nd, Fang R, Moore RJ, Smith RD, Xiao W, Davis RW, Tompkins RG: **Ultra-high-efficiency strong cation exchange LC/RPLC/MS/MS for high dynamic range characterization of the human plasma proteome**. *Analytical chemistry* 2004, **76**(4):1134-1144.

76. von Muhlen CA, Tan EM: **Autoantibodies in the diagnosis of systemic rheumatic diseases**. *Seminars in arthritis and rheumatism* 1995, **24**(5):323-358.

77. Kanter JL, Narayana S, Ho PP, Catz I, Warren KG, Sobel RA, Steinman L, Robinson WH: **Lipid microarrays identify key mediators of autoimmune brain inflammation**. *Nat Med* 2006, **12**(1):138-143.

78. Robinson WH, DiGennaro C, Hueber W, Haab BB, Kamachi M, Dean EJ, Fournel S, Fong D, Genovese MC, de Vegvar HE *et al*: **Autoantigen microarrays for multiplex characterization of autoantibody responses**. *Nat Med* 2002, **8**(3):295-301.

79. Quintana FJ, Hagedorn PH, Elizur G, Merbl Y, Domany E, Cohen IR: **Functional immunomics: microarray analysis of IgG autoantibody repertoires predicts the future response of mice to induced diabetes**. *Proc Natl Acad Sci U S A* 2004, **101 Suppl 2**:14615-14621.

80. Bertone P, Snyder M: **Advances in functional protein microarray technology**. *The FEBS journal* 2005, **272**(21):5400-5411.

81. Lueking A, Possling A, Huber O, Beveridge A, Horn M, Eickhoff H, Schuchardt J, Lehrach H, Cahill DJ: **A nonredundant human protein chip for antibody screening and serum profiling**. *Mol Cell Proteomics* 2003, **2**(12):1342-1349.

82. Hutchens TW, Yip TT: **New desorption strategies for the mass spectrometric analysis of macromolecules**. *Rapid Commun Mass Spectrom* 1993, **7**:576-580.

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

84. Petricoin EF, 3rd, Ornstein DK, Paweletz CP, Ardekani A, Hackett PS, Hitt BA, Velassco A, Trucco C, Wiegand L, Wood K *et al*: **Serum proteomic patterns for detection of prostate cancer**. *Journal of the National Cancer Institute* 2002, **94**(20):1576-1578.

85. Ornstein DK, Rayford W, Fusaro VA, Conrads TP, Ross SJ, Hitt BA, Wiggins WW, Veenstra TD, Liotta LA, Petricoin EF, 3rd: **Serum proteomic profiling can discriminate prostate cancer from benign prostates in men with total prostate specific antigen levels between 2.5 and 15.0 ng/ml**. *J Urol* 2004, **172**(4 Pt 1):1302-1305.

86. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC *et al*: **Use of proteomic patterns in serum to identify ovarian cancer**. *Lancet* 2002, **359**(9306):572-577.

87. Kozak KR, Amneus MW, Pusey SM, Su F, Luong MN, Luong SA, Reddy ST, Farias-Eisner R: **Identification of biomarkers for ovarian cancer using strong anion-exchange ProteinChips: potential use in diagnosis and prognosis**. *Proc Natl Acad Sci U S A* 2003, **100**(21):12343-12348.

88. Bhattacharyya S, Siegel ER, Petersen GM, Chari ST, Suva LJ, Haun RS: **Diagnosis of pancreatic cancer using serum proteomic profiling**. *Neoplasia* 2004, **6**(5):674-686.

89. Maurya P, Meleady P, Dowling P, Clynes M: **Proteomic approaches for serum biomarker discovery in cancer**. *Anticancer Res* 2007, **27**(3A):1247-1255.

90. Petricoin EF, Zoon KC, Kohn EC, Barrett JC, Liotta LA: **Clinical proteomics: translating benchside promise into bedside reality**. *Nat Rev Drug Discov* 2002, **1**(9):683-695.

91. Winer S, Tsui H, Lau A, Song A, Li X, Cheung RK, Sampson A, Afifiyan F, Elford A, Jackowski G *et al*: **Autoimmune islet destruction in spontaneous type 1 diabetes is not beta-cell exclusive**. *Nat Med* 2003, **9**(2):198-205.

92. Saouda M, Romer T, Boyle MD: **Application of immuno-mass spectrometry to analysis of a bacterial virulence factor**. *Biotechniques* 2002, **32**(4):916, 918, 920, 922-913.

93. Beher D, Wrigley JD, Owens AP, Shearman MS: **Generation of C-terminally truncated amyloid-beta peptides is dependent on gamma-secretase activity**. *J Neurochem* 2002, **82**(3):563-575.

94. Carrette O, Demalte I, Scherl A, Yalkinoglu O, Corthals G, Burkhard P, Hochstrasser DF, Sanchez JC: **A panel of cerebrospinal fluid potential biomarkers for the diagnosis of Alzheimer's disease**. *Proteomics* 2003, **3**(8):1486-1494.

95. Arroyo CM, Broomfield CA, Hackley BE, Jr.: **The role of interleukin-6 (IL-6) in human sulfur mustard (HD) toxicology**. *Int J Toxicol* 2001, **20**(5):281-296.

96. Norwitz ER, Tsen LC, Park JS, Fitzpatrick PA, Dorfman DM, Saade GR, Buhimschi CS, Buhimschi IA: **Discriminatory proteomic biomarker analysis identifies free hemoglobin in the cerebrospinal fluid of women with severe preeclampsia**. *Am J Obstet Gynecol* 2005, **193**(3 Pt 2):957-964.

97. Ruetschi U, Zetterberg H, Podust VN, Gottfries J, Li S, Hviid Simonsen A, McGuire J, Karlsson M, Rymo L, Davies H *et al*: **Identification of CSF biomarkers for frontotemporal dementia using SELDI-TOF**. *Exp Neurol* 2005, **196**(2):273-281.

98. Villar-Garea A, Griese M, Imhof A: **Biomarker discovery from body fluids using mass spectrometry**. *J Chromatogr B Analyt Technol Biomed Life Sci* 2007, **849**(1-2):105-114.

99. Hsu PI, Chen CH, Hsieh CS, Chang WC, Lai KH, Lo GH, Hsu PN, Tsay FW, Chen YS, Hsiao M *et al*: **Alpha1-antitrypsin precursor in gastric juice is a novel biomarker for gastric cancer and ulcer**. *Clin Cancer Res* 2007, **13**(3):876-883.

100. Tomosugi N, Kitagawa K, Takahashi N, Sugai S, Ishikawa I: **Diagnostic potential of tear proteomic patterns in Sjogren's syndrome**. *J Proteome Res* 2005, **4**(3):820-825.

101. Ryu OH, Atkinson JC, Hoehn GT, Illei GG, Hart TC: **Identification of parotid salivary biomarkers in Sjogren's syndrome by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry and two-dimensional difference gel electrophoresis**. *Rheumatology (Oxford)* 2006, **45**(9):1077-1086.

102. Uchida T, Fukawa A, Uchida M, Fujita K, Saito K: **Application of a novel protein biochip technology for detection and identification of rheumatoid arthritis biomarkers in synovial fluid**. *J Proteome Res* 2002, **1**(6):495-499.

103. de Seny D, Fillet M, Meuwis MA, Geurts P, Lutteri L, Ribbens C, Bours V, Wehenkel L, Piette J, Malaise M *et al*: **Discovery of new rheumatoid arthritis biomarkers using the surface-enhanced laser desorption/ionization time-of-flight mass spectrometry ProteinChip approach**. *Arthritis Rheum* 2005, **52**(12):3801-3812.

104. Lewczuk P, Esselmann H, Meyer M, Wollscheid V, Neumann M, Otto M, Maler JM, Ruther E, Kornhuber J, Wiltfang J: **The amyloid-beta (Abeta) peptide pattern in cerebrospinal fluid in Alzheimer's disease: evidence of a novel carboxyterminally elongated Abeta peptide**. *Rapid Commun Mass Spectrom* 2003, **17**(12):1291-1296.

105. Check E: **Proteomics and cancer: running before we can walk?** *Nature* 2004, **429**(6991):496-497.

106. Baggerly KA, Morris JS, Coombes KR: **Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments**. *Bioinformatics (Oxford, England)* 2004, **20**(5):777-785.

107. Baggerly KA, Morris JS, Edmonson SR, Coombes KR: **Signal in noise: evaluating reported reproducibility of serum proteomic tests for ovarian cancer**. *Journal of the National Cancer Institute* 2005, **97**(4):307-309.

108. Sorace JM, Zhan M: **A data review and re-assessment of ovarian cancer serum proteomic profiling**. *BMC bioinformatics* 2003, **4**:24.

109. Diamandis EP: **Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations**. *Mol Cell Proteomics* 2004, **3**(4):367-378.

110. Diamandis EP: **Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems**. *Journal of the National Cancer Institute* 2004, **96**(5):353-356.

111. Diamandis EP: **Peptidomics for cancer diagnosis: present and future**. *Journal of proteome research* 2006, **5**(9):2079-2082.

112. Ransohoff DF: **Bias as a threat to the validity of cancer molecular-marker research**. *Nature reviews* 2005, **5**(2):142-149.

113. Mischak H, Apweiler R, Banks RE, Conaway M, Coon J, Dominiczak A, Ehrich JHH, Fliser D, Girolami M, Hermjakob H *et al*: **Clinical proteomics: A need to define the field and to begin to set adequate standards**. *Proteomics Clinical Applications* 2007, **1**(2):148-156.

114. Fisher WG, Rosenblatt KP, Fishman DA, Whiteley GR, Mikulskis A, Kuzdzal SA, Lopez MF, Tan NC, German DC, Garner HR: **A robust biomarker discovery pipeline for high-performance mass spectrometry data**. *Journal of bioinformatics and computational biology* 2007, **5**(5):1023-1045.

115. Hu J, Coombes KR, Morris JS, Baggerly KA: **The importance of experimental design in proteomic mass spectrometry experiments: some cautionary tales**. *Briefings in functional genomics & proteomics* 2005, **3**(4):322-331.

116. Banks RE, Stanley AJ, Cairns DA, Barrett JH, Clarke P, Thompson D, Selby PJ: **Influences of blood sample processing on low-molecular-weight proteome identified by surface-enhanced laser desorption/ionization mass spectrometry**. *Clinical chemistry* 2005, **51**(9):1637-1649.

117. Drake SK, Bowen RA, Remaley AT, Hortin GL: **Potential interferences from blood collection tubes in mass spectrometric analyses of serum polypeptides**. *Clinical chemistry* 2004, **50**(12):2398-2401.

118. Hsieh SY, Chen RK, Pan YH, Lee HL: **Systematical evaluation of the effects of sample collection procedures on low-molecular-weight serum/plasma proteome profiling**. *Proteomics* 2006, **6**(10):3189-3198.

119. Timms JF, Arslan-Low E, Gentry-Maharaj A, Luo Z, T'Jampens D, Podust VN, Ford J, Fung ET, Gammerman A, Jacobs I *et al*: **Preanalytic influence of sample handling on SELDI-TOF serum protein profiles**. *Clinical chemistry* 2007, **53**(4):645-656.

120. Rai AJ, Gelfand CA, Haywood BC, Warunek DJ, Yi J, Schuchard MD, Mehigh RJ, Cockrill SL, Scott GB, Tammen H *et al*: **HUPO Plasma Proteome Project specimen collection and handling: towards the standardization of parameters for plasma proteome samples**. *Proteomics* 2005, **5**(13):3262-3277.

CHAPTER TWO
Methodology

2.1 HIGH-THROUGHPUT PLATFORM DEVELOPMENT

Presently, there does not exist a technology or experimental approach that can simultaneously address all the challenges listed for proteomics biomarker discovery in Section 1.3.1. Even though SELDI had demonstrated potential in disease classification, as with any immature technology, all potential sources of bias will not be apparent until it has undergone objective evaluation over time by different users. Since a seminal report in 2002 [1], technological advancements in MS instrumentation and data processing methods have allowed ongoing improvements to the platform [2]. In light of all these changes, the underlying principle of this technology remains unchanged: to utilize mass spectra generated from two different sample groups to perform comparative data mining by coupling MS data to heuristic pattern recognition and data mining algorithms for discovery of differential diagnostic peaks (Fig. 2.1). As such, the fidelity of the mass peaks produced is of utmost importance.

The ideal proteomic platform for disease biomarker discovery should be able to: (i) analyze thousands of proteins in parallel in a high-throughput fashion using minuscule samples that are readily available, (ii) simplify the proteome dramatically to confer the resolution and sensitivity necessary to probe for low abundance proteins and to approximate the dynamic range of current analytical tools and (iii) be fully automated to ensure reproducibility without being labor intensive. The subsequent adoption of sophisticated bioinformatics tools for rigorous data analysis to identify differential species should result in a panel of robust biomarkers that are disease specific.

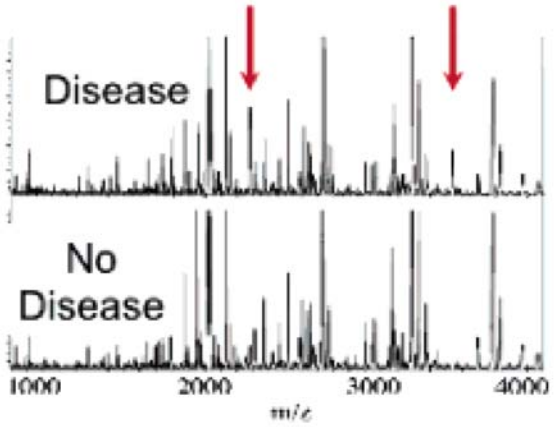


Figure 2.1 **Comparative proteomics for biomarker discovery**. The mass spectra from two sample groups (disease and non-disease shown here) are compared via bioinformatics to seek diagnostic marker peaks. Two differential peaks in the disease group are indicated by arrows.

This chapter describes an undertaking to evaluate all possible variables that are within our control in the SELDI-based biomarker discovery workflow, from sample processing to candidate marker identification. The workflow is divided into three main modules: (i) sample processing, (ii) data acquisition and (iii) data analysis. The primary objective is the development and implementation of a methodology to serve as a robust front-end biomarker discovery tool that will facilitate subsequent identification and verification efforts of candidate disease markers. To this end, the discovery process entails the unbiased binary comparison between disease and control samples, controlling for noise from non-disease related conditions. The immediate product of this phase is a list of mass peaks found to be differential between the two states compared based on semiquantitative assessment of their relative protein/peptide abundance in the MS data. The tens or hundreds of initial candidates will undergo stringent screening and be reduced to a smaller set of peaks that possess true discriminatory power. The second objective is to identify these peaks and perform preliminary verification on them. Molecular identification of

the putative biomarker is critical if one wishes to eventually develop a more conventional, non-MS-based assay for the marker, such as an ELISA. This general biomarker discovery process is depicted in Figure 2.2.

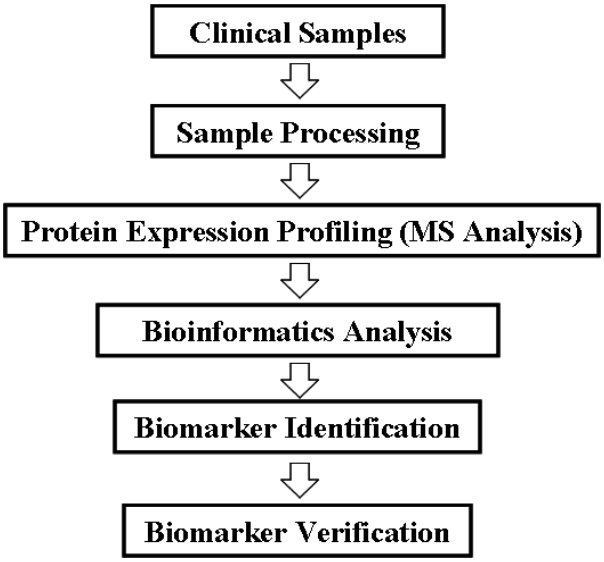


Figure 2.2 **General biomarker discovery process**. Preanalytical variables are addressed within Clinical Samples, analytical variables in Sample Processing and Protein Expression Profiling and postanalytical variables in Bioinformatics Analysis.

The hierarchical structure of the workflow is shown in Figure 2.3. At the top of the hierarchy is the choice of the samples to be compared, which as mentioned previously would ideally control for all biological variables other than the presence or absence of disease. However, this is extremely difficult to achieve. The middle level of the hierarchy assays variation between the samples within a given group, capturing the major source of biological variation. Biological variations are from environmental and genetic origins in a large heterogeneous population like humans. The lowest level of hierarchy involves replicate runs from the same sample, and captures the inherent analytical variation [3].

The importance and optimization of the parameters involved throughout the workflow are detailed below.

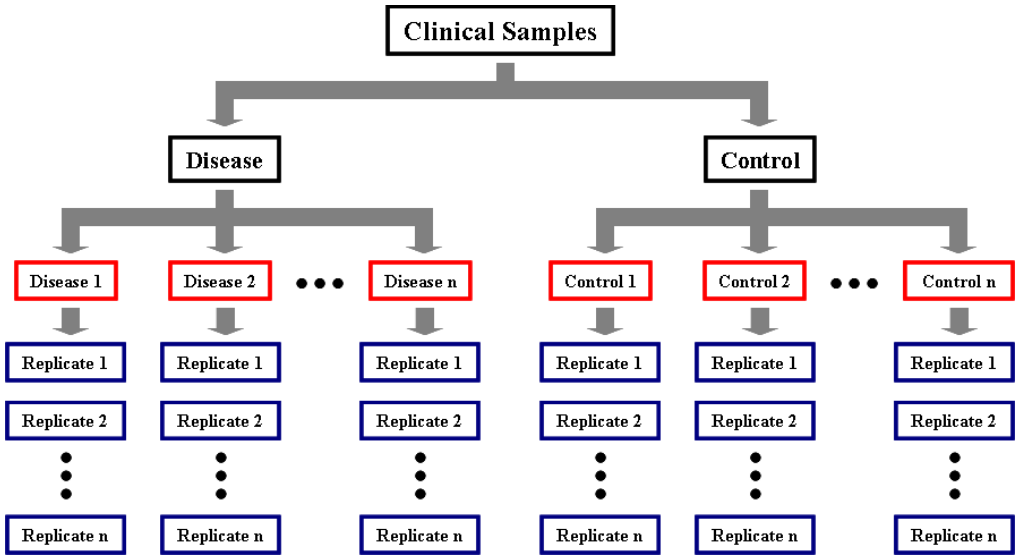


Figure 2.3 Experimental design to reduce biological (red) and analytical (blue) variations**.**

2.2 MODULE I: SAMPLE PROCESSING

This section encompasses discussion on sample source, sample pre-treatment conditions, sample dilution factors, and sample fractionation. Different sample processing conditions were evaluated as the sensitivity of the assay is largely dependent on the sample preparation than mass spectrometric methods. An optimized condition for the parameters in Modules I and II is defined as the one that provides the greatest number of peaks in the mass spectrum. The reasoning is that the larger the proportion of total human proteome detected, the greater the chance of finding proteins/peptides that are affected by the disease.

2.2.1 Sample source

Biological samples that are amenable for proteomics biomarker discovery include cells, tissues, and bodily fluids. Practical and technical considerations ultimately dictate the preference of one source of sample over the others. Cell and tissue lysates, though applicable to this workflow, involve cell disruption variations that are difficult to control and represents an additional source of sample preparation bias. Intact tissues demand the invasive procedure of biopsy and are predominantly used in MS imaging experiments, such as laser capture microdissection (LCM) MS experiments [4, 5]. Therefore, only readily accessible biological fluids will be considered here as they involve minimal risk and cost to obtain.

Blood tests are the gold standard of clinical diagnosis, making blood (plasma and serum) a logical fluid to use for biomarker discovery. By definition, serum is the undiluted, extracellular portion of blood after adequate coagulation is complete. Plasma is the virtually cell-free supernatant of blood containing anticoagulant obtained after centrifugation [6]. Plasma introduces too many preanalytical variables during collection as a result of the anticoagulants used [6]. Therefore, serum is more desirable, in addition to also being the most archived specimen [7]. An added advantage of serum is proteome simplification through the removal of fibrinogen, one of the top five most abundant proteins in plasma, during coagulation. In the narcolepsy study (Chapter 4), serum samples were obtained from the Center for Narcolepsy at Stanford University.

Disease-specific markers that arise locally tend to experience diminishing signal with distance from the affected site through dilution post-leakage into the circulation. Hence, analyzing proximal biological fluids either close to or in direct contact with the disease site may enhance biomarkers concentration [8]. For example, in a study of 33 patients with ovarian cancer, median CA-125 levels (U/ml) were 696 in serum, 18,563 in ascites and 44,850 in cyst fluid [9]. Furthermore, the initial protein biomarker discovery in proximal fluid may be a surrogate for its availability in systemic circulation, if a blood test is the eventual goal. In the multiple sclerosis

study (Chapter 3), CSF samples were obtained from clinical CSF banks. CSF was chosen because it is in direct contact with the extracellular space in the brain, and hence diseases related to the central nervous system (CNS) such as multiple sclerosis will potentially affect the biochemical composition of this body fluid. In addition, CSF has a lower dynamic range in protein concentration (10^8) than plasma (10^{10}) (Fig. 2.4) and this in itself is a simplification of the proteome to be analyzed.

Most proteomics discovery efforts are conducted with biological materials selected to maximize the detection of meaningful protein differences while minimizing the sample number required for analysis in the interest of throughput. If the sample number is kept small (<10), the observed differences between the two sets of specimen are in danger of being overinterpreted when extrapolated to the generalized population, known as the problem of sparse data [10]. There is currently no consensus on the ideal minimum number of samples required for biomarker discovery efforts although a reasonable representative selection of marker candidates can be achieved from a minimum of 15 samples [11]. As a pilot study, a total of 30 samples were used in narcolepsy. The larger multiple sclerosis study involved 60 samples.

Our sample sets are representative of the target population as they originate from disease centers that collect these samples routinely for diagnosis. Both studies are described in detail in Chapter 3 (Case Study I: Multiple Sclerosis) and Chapter 4 (Case Study II: Narcolepsy).

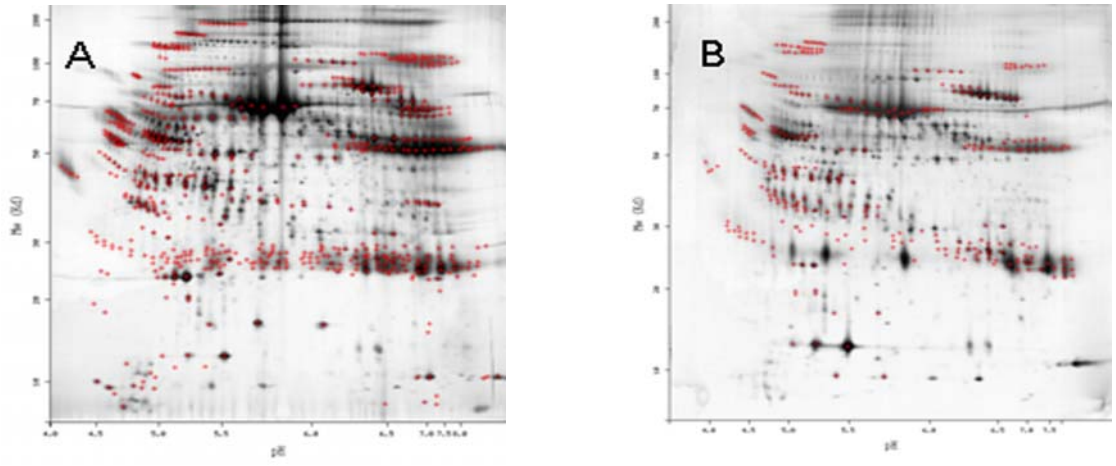


Figure 2.4 **Relative comparison of proteome complexity of human serum (A) and cerebrospinal fluid (B) in 2DGE**. Even though there are less proteins in CSF, the dynamic range between the high and low abundance proteins still exists. Molecular weight is on the y-axis, isoelectric point (pI) on the x-axis.

2.2.2 Sample pre-treatment

Pre-treating complex biological samples with urea to disrupt protein-protein interactions had been suggested to enhance the number of detectable mass peaks in protein profiling experiments [12]. This is also an appealing option as it will render the samples more compatible with downstream biomarker enrichment technologies like reverse phase HPLC and the two-dimensional protein separation platform (ProteomeLab PF2D, Beckman Coulter) that denature samples prior to separation. This sample pre-treatment option was evaluated by comparing the mass spectra from a non-treated, native serum sample and a urea-treated serum sample (Fig. 2.5) on IMAC30 ProteinChip arrays pre-charged with nickel. The sample pre-treatment condition that produced the

greater number of peaks, noticeably in the LMW mass range (<2,500 Da), was the native condition. Therefore, all serum samples in our study were analyzed in their native form.

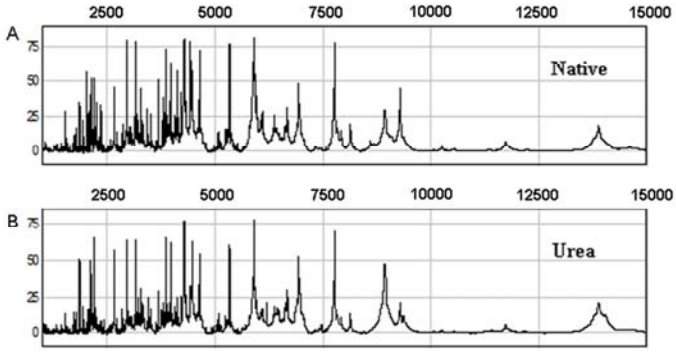


Figure 2.5 **Optimization of sample pre-treatment condition**. Mass spectra correspond to the same standard serum sample analyzed either in the non-treated, native form (A) or after urea pre-treatment (B). More peaks were observed in the LMW region of the native sample. Signal intensity (y-axis) is plotted against the *m/z* ratio (x-axis).

2.2.3 Sample dilution factor

The optimization of the sample dilution ratio when introduced to the ProteinChip arrays is a necessary undertaking for every SELDI-based proteomic biomarker discovery study. This is essential as the protein composition and concentration vary with biological sample and as such, affect the optimal dilution factor that will maximize binding of proteins/peptides to the capture surface and minimize ion suppression during MS analysis (as evaluated by the number of peaks per mass spectrum). To demonstrate, a standard serum sample was incubated on the same chip either neat or diluted 10-fold. As apparent from Figure 2.6, even though the protein concentration is higher in the neat sample, the 10-fold diluted sample provides an increase in the number of observed mass peaks. The optimization of CSF and albumin-bound subproteome samples were

carried out for the multiple sclerosis and narcolepsy studies, as detailed in their respective chapters.

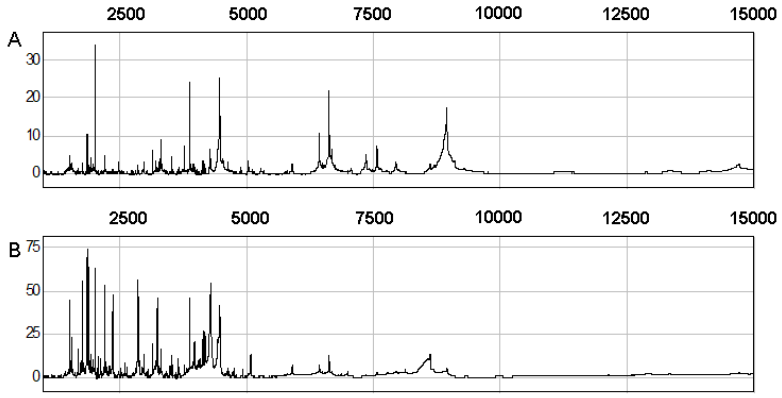


Figure 2.6 **Optimization of sample dilution factor**. Mass spectra correspond to the same standard serum sample analyzed either neat (A) or 10-fold diluted (B) on an IMAC30 chip.

2.2.4 Sample fractionation

As is often the case, the overwhelming presence of peptides derived from the most abundant proteins causes significant ion suppression of lower abundance analytes. Therefore, pre-fractionation is an absolute prerequisite if disease biomarkers in the ng/ml concentration range are to be detected. A simplified proteome will also reduce the competition among the protein constituents for the limited binding sites on the chip surface.

The most expedient way to gain up to two orders of magnitude in proteome coverage is by depleting the most abundant-proteins [13-15]. In addition, a number of other pre-fractionation tools that either target subproteomes (glyco- or phosphoproteomes[16], organelle subproteomes) or separate based on biochemical properties (chromatography based on ion charges or metal binding ability, isoelectric focusing separation) are used commonly. It is worth noting that all

sample processing steps and fractionation methods result in some level of analyte loss [17] and could introduce variability between samples.

In our studies with CSF, the need for pre-fractionation was less stringent. As the protein concentration of CSF is already two orders of magnitude lower than that of plasma, no pre-fractionation step was incorporated in the workflow. On the other hand, complex serum samples were fractionated into albumin-enriched and albumin-depleted fractions. Only the albumin-enriched fraction was subject to analysis, effectively simplifying the serum proteome to just the albumin-bound cargo. A more elaborate fractionation of serum was not performed as a compromise between in-depth analysis and throughput. In spite of all the automation, the goal is to limit sample processing steps to where a statistical number of samples can be analyzed within a reasonable period of time.

2.3 MODULE II: DATA ACQUISITION

This section discusses matrix optimization, surface chemistry optimization, and technical reproducibility. In addition to the sample processing steps covered in Module I, a huge determinant of the resultant mass spectra quality is data acquisition conditions during the analytical phase. The identified sources of analytical and physical variability are evaluated and optimized as discussed below.

2.3.1 Matrix optimization

As with MALDI, the spectral quality and profile obtained from SELDI TOF MS is dependent on the specific type of matrix used. CHCA is generally favored for the detection of small peptides, 3,5-dimethoxy-4-hydroxycinnamic / sinapanic acid (SA) for small and medium-sized proteins, while heavily glycosylated and large proteins are detected more easily using 2,5-dihydroxybenzoic acid (DHB) or ferulic acid matrix [18]. We evaluated the two most common

matrices in MS, namely CHCA and SA, at different concentrations to determine the best condition for optimal display of spectrum peaks. CHCA at 5 mg/ml provided the most number of peaks with considerable intensity at the 1,000 to 10,000 *m/z* mass range (Fig. 2.7). CHCA has also been shown to be less amenable to in-source decay artifacts in SELDI [19]. CHCA (optional recrystallization) from different vendors was evaluated and that from LaserBio Labs (France) provided the best quality and most reproducible spectra without the need for recrystallization. It should be noted that cluster formation is notorious with CHCA. These clusters are usually observed as sodium and potassium adducts, complicating the mass spectrum in the 500 – 1,300 mass range [20-22]. As such, only peaks greater than 1,000 Da are subject to data analysis and of all the peaks that are listed as statistically differential, only those that are beyond this mass range are given priority in identification and verification stages.

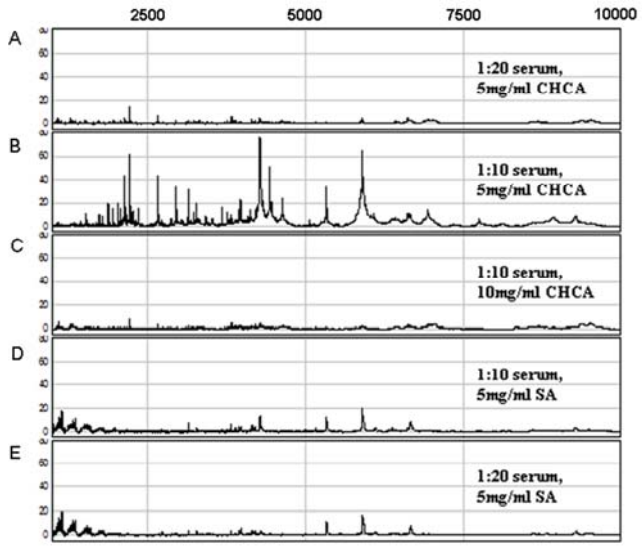
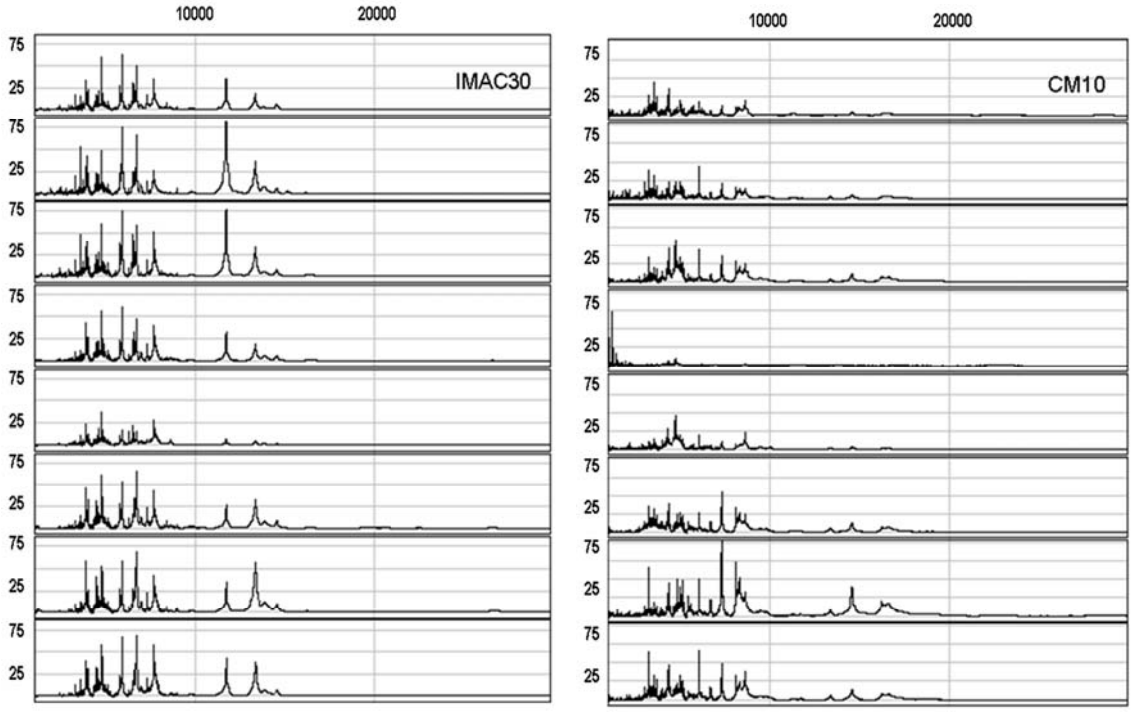


Figure 2.7 **Optimization of serum dilution factor and ionizing matrix concentration**. Shown are mass spectra corresponding to the same serum sample diluted 10- and 20-fold and evaluated with either CHCA (A, B) or SA (D, E). CHCA was also evaluated at different concentrations, 5 mg/ml (A, B) and 10 mg/ml (C).

2.3.2 Surface chemistry optimization

The strength of SELDI lies in combining separation platforms and MS capabilities to expand the portion of complex biological samples that can be profiled. The chromatographic surface on the protein chips confers another level of proteome simplification by selectively binding only a subset of the proteins in the sample. The larger the number of proteins is in this subset, the greater the chance of the biomarker of interest is captured. Four main capture surface chemistries were evaluated for optimal peak observation: cation exchange (CM10), anion exchange (Q10), reverse phase (H50), and metal affinity (IMAC30). A set of eight patient serum samples was run on chips with the aforementioned surface chemistries.



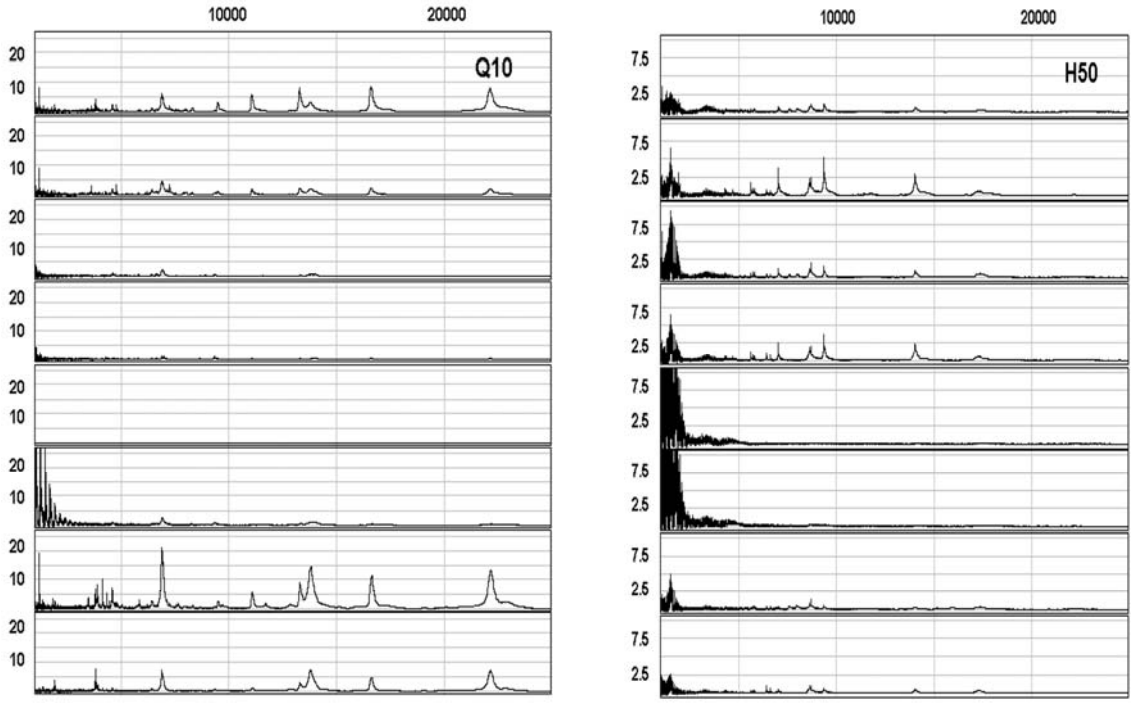


Figure 2.8 **Optimization of surface chemistry**. Shown are IMAC30, CM10, Q10, and H50 surfaces evaluated for peak production with the same set of eight different patient samples.

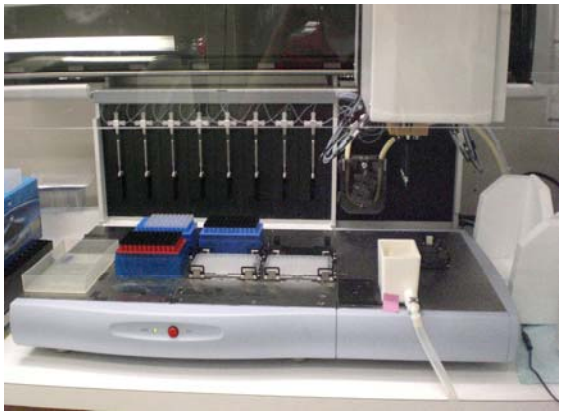
As is apparent to the eye, IMAC30 proved to be the best peak-producing platform, followed by CM10 (Fig. 2.8). Thus, IMAC30 was the retentate surface of choice in the workflow. Less in-source decay artifacts were also observed with the IMAC30 chips [19]. IMAC30 was further optimized using different charging metal ions such as nickel, copper, and iron. This is necessary as transition metals have idiosyncratic binding properties that result in differing peptide/protein profiles that are produced. Even though the manufacturer's protocol recommended copper as the metal ion of choice, we found IMAC30 chips charged with nickel(II) ions displayed more output peaks, in agreement with others [12].

2.3.3 Technical reproducibility

The issue of reproducibility [23-27] is an inherent problem with discovery-based research, as echoed by microarray gene expression analyses [28]. Peak intensities with SELDI TOF MS are highly sensitive to experimental details. Day-to-day, lot-to-lot, and machine-to-machine variances resulting from sample handling and storage are all too familiar to SELDI users [29-30]. However, reproducibility is attainable through proper experimental design and incorporation of quality control checks throughout the process [31].

Even though the reproducibility on the ProteinChip arrays have significantly increased upon automation of the manufacturing process, lot-to-lot variability still exists and was experienced first hand. Therefore, all the chips to be used within the same study were selected from the same batch and a chip from that batch is first evaluated with serum for spectral quality as part of the quality control of the workflow. A pooled reference standard serum sample obtained from the National Institute of Standards and Technology (NIST) is religiously applied as the positive control in every study to assess the overall integrity of the workflow from sample processing to the operating condition of the mass spectrometer.

Operator bias in sample preparation will also introduce artifacts in the output spectrum. This is minimized by automating the whole workflow as much as possible with minimum human intervention. The PerkinElmer MultiPROBE II PLUS HT EX liquid handler was incorporated into the workflow to process up to twelve ProteinChip arrays (96 spots) in parallel using the ProteinChip array bioprocessor (Ciphergen) for high-throughput analysis (Fig. 2.9). Automation with a liquid handler for sample deposition, washing steps, and matrix deposition is crucial to achieve accurate reproducibility, minimize process-driven variability, and increase robustness of the platform [31].



A

B

Figure 2.9 **High-throughput ProteinChip analysis**. (A) Automation in sample preparation for SELDI analysis. Shown here is the MultiPROBE II liquid handler from PerkinElmer. (B) High-throughput handling of ProteinChip arrays. Shown here is the bioprocessor from Ciphergen that can handle up to 12 arrays simultaneously.

An in depth understanding of mass spectrometer operation is essential to capitalize on its strengths and know its limitations when optimizing acquisition parameters. Relatively small changes to the operating conditions can be amplified to produce fairly large differences in mass spectra, making it difficult to maintain consistent, reproducible results. Semmes *et al.* [31] have shown that the performance of the SELDI TOF MS instrument from Ciphergen may change over time because of varying laser intensity and detector sensitivity. Therefore, in order to minimize day-to-day variability seen in signal drift, all samples in the studies described in this dissertation were run in one setting to hold all technical variables constant. The laser intensity was set based on the condition that provided the best spectrum with the NIST serum sample. Prior to the high-throughput run, the instrument was calibrated with an external calibrant to ensure mass accuracy. Ultimately, the performance of the mass spectrometer employed for mass peak production is the major determinant of peak fidelity as discussed in the following section.

2.3.3.1 Mass spectrometer performance comparison

The strength of MS protein profiling is not in direct protein identification but rather in linking the signal intensity of a collection of protein peaks to a clinical outcome using bioinformatics. The amount of information that can be extracted with high confidence from a complex sample is a function of the sensitivity, resolution, and mass accuracy of the instrumentation. The values for both *m/z* and peak intensity are critical components in the construction of the SELDI spectral profile. For differential marker peak discovery, the mass spectra from the two sample groups are compared semiquantitatively at every detected peak for statistically significant changes (Fig. 2.10). Minimal mass drift is important for accurate peak-to-peak comparison during data analysis. In this respect, the reliability and reproducibility of the chip and measurement system are of paramount importance.

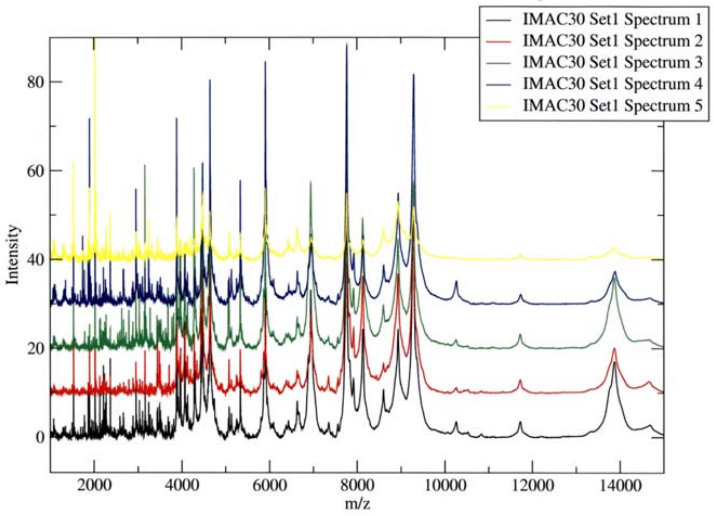
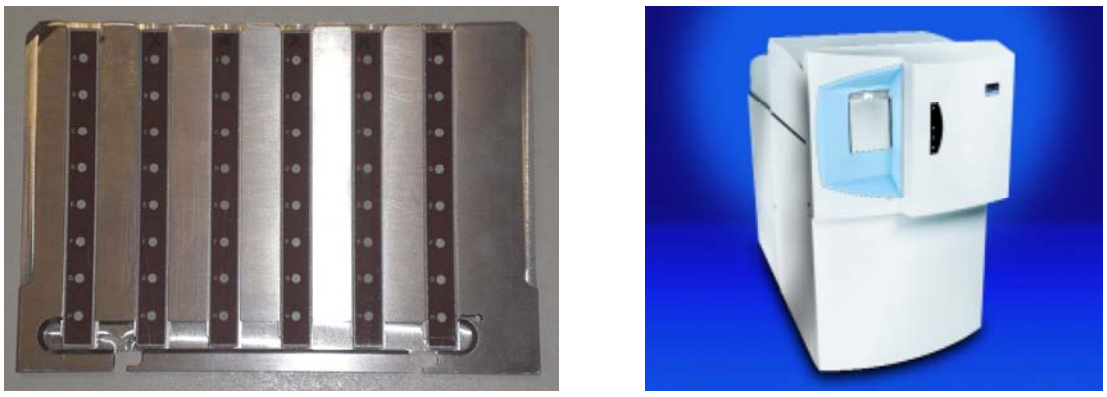


Figure 2.10 **Peak comparison across mass spectra to uncover differential mass peaks**. In this simplified view, spectra from five different samples (Spectrum 1 to Spectrum 5) are compared at every peak in the search for statistical difference in intensity, highlighting the importance of peak alignment.

Reproducibility of MS-based profiling studies can be enhanced by improving instrument design. The incorporation of high resolution mass spectrometers to the SELDI platform has been shown to yield superior diagnostic profiles to those from low-resolution instruments in terms of sensitivity and specificity as a result of both the increased number of peaks seen and much better reproducibility [29]. Therefore, the performance of the two ProteinChip-compatible mass spectrometers that were available to us for profiling studies was compared. In this study, 90 sample spots were run in parallel and spectra were acquired from the same ProteinChip spots on the two instruments. This is accomplished by first acquiring data on the PerkinElmer prOTOF 2000 MALDI O-TOF mass spectrometer [32] by setting the laser to ablate in a circular pattern on the spot. ProteinChip arrays were placed in a custom made adapter for mass spectrometry analysis in the prOTOF mass spectrometer (PerkinElmer/SCIEX) (Fig. 2.11). Its orthogonal design enabled a single external mass calibrant to achieve better than 5 ppm mass accuracy over the 1,000 to 10,000 mass range acquired. A 2-point external calibration of the prOTOF instrument was performed before spectra acquisition in a batch mode, four arrays at a time. The prOTOF data files generated an average of 1 million data points per spectrum.



A B

Figure 2.11 **High resolution mass spectrometry analysis**. (A) Custom made adapter for ProteinChip analysis on prOTOF. (B) The prOTOF mass spectrometer from PerkinElmer.

The same ProteinChip arrays were removed from prOTOF and loaded into the Ciphergen PBS-IIc mass spectrometer. Data acquisition was performed through ablation of the same spots in a linear fashion (Fig. 2.12). Acquisition was performed in a batch mode of 12 arrays in the cassette that accompanied the instrument. Calibration was performed externally with the All-in-1 Peptide Calibrant (Ciphergen) with a laser intensity of 170 and a sensitivity of 9. The PBS-IIc data files generated an average of 40,000 data points per spectrum.

A B

Figure 2.12 **Low resolution mass spectrometry analysis**. (A) Cassette for high-throughput analysis on PBS-IIc. (B) The PBS-IIc mass spectrometer from Ciphergen.

To determine the mass accuracy of the instruments, the mass of a prominent peak at *m/z* = 2021 across all spectra was examined. The mass accuracy of the prOTOF was found to be around 5 ppm whereas that of PBS-IIc was around 1,000 ppm (Panels A and B, Fig. 2.13). Peak misalignment is also apparent visually when another peak at *m/z* = 3805 was examined across all spectra (Panels C and D, Fig. 2.13). In addition, there was an observed significant mass drift (0.1-0.2%) for the same peak intra- and interchips (Panel E, Fig. 2.13), in agreement with others [33].

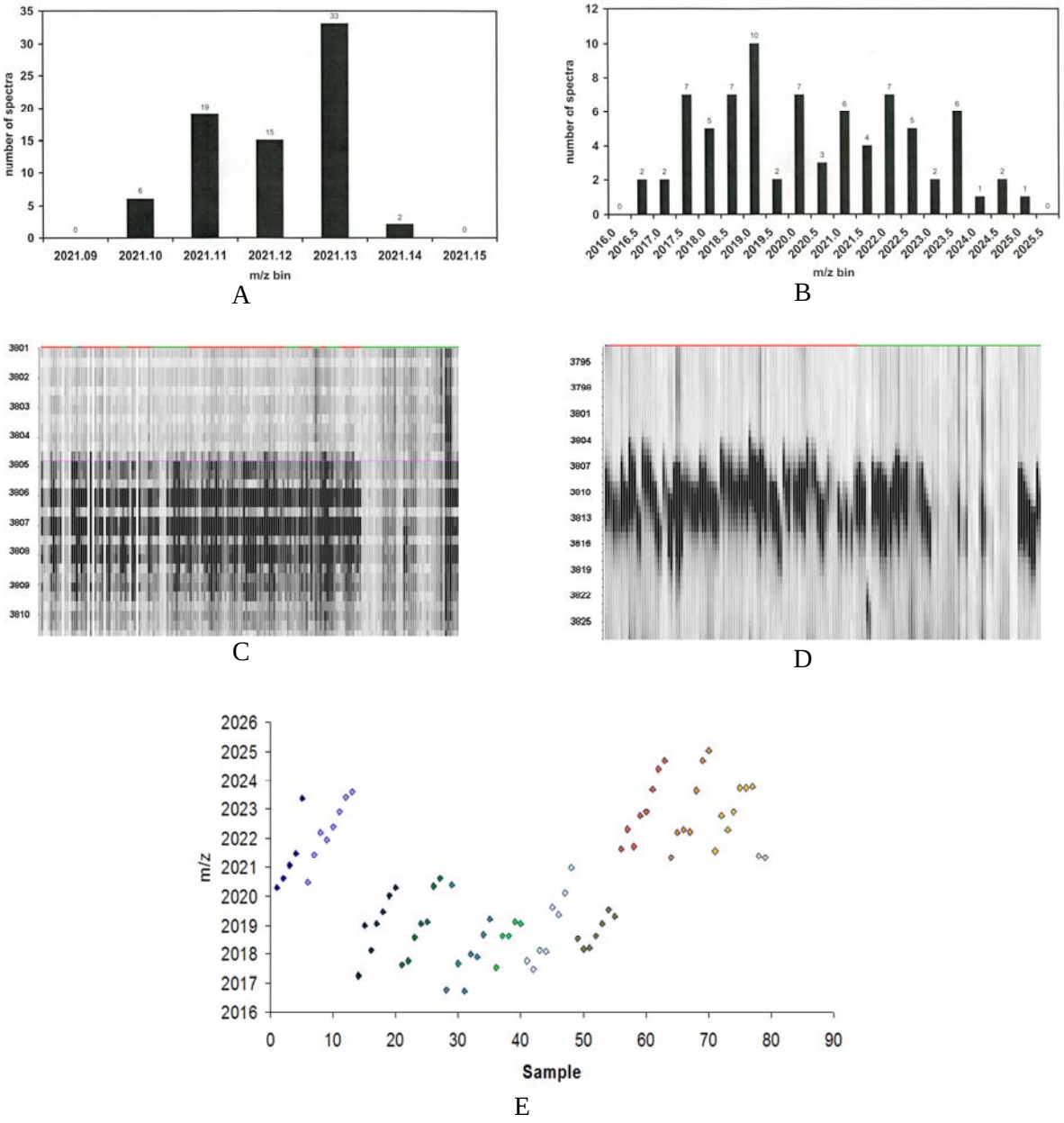


Figure 2.13 **Performance comparison between the prOTOF and PBS-IIc mass spectrometers**. (A and B) The mass accuracy for a selected peak at *m/z* = 2021 was evaluated across spectra obtained from both mass spectrometers. The mass accuracy was determined to be 5 ppm for prOTOF (Panel A) and 1,000 ppm for PBS-IIc (Panel B). The number of spectra corresponding to each mass value (y-axis) was plotted against the mass range (x-axis). (C and D)

When a peak at *m/z* = 3805 was examined, significant mass drift was observed across spectra in PBS-IIc (Panel D) and was almost non-existent in prOTOF (Panel C). The mass range (y-axis) was plotted against all acquired spectra (x-axis). (E) The mass drift of peak 2021 is shown intra- and interchips on the PBS-IIc system. Mass variation was observed for the same peak for samples within the same chip (same color) and between chips (different color). The mass range (y-axis) was plotted against the sample run order (x-axis).

An alignment strategy was developed in-house (Fig. 2.14) in collaboration with the Garner Laboratory (UT Southwestern) to address the mass drift problem evident in the PBS-IIc spectra. The purpose of this initiative was to determine if the peak alignment of the PBS-IIc data could be improved and to evaluate how the mass accuracy of the aligned data compares to the prOTOF spectra. From within the PBS-IIc data, the spectrum with the most number of peaks was chosen as the reference to which the remaining spectra were aligned. Peaks with undetermined positions were extrapolated from neighboring peaks. Table 2.1 shows the mass accuracy across six peaks that span the acquired mass range for the raw and aligned data from PBS-IIc as compared to the raw prOTOF data. Even though significantly improved (Fig. 2.15), the aligned PBS-IIc data do not fair as well as the raw prOTOF data, with mass accuracy still in the hundreds ppm compared to <10 ppm attainable in the prOTOF data.

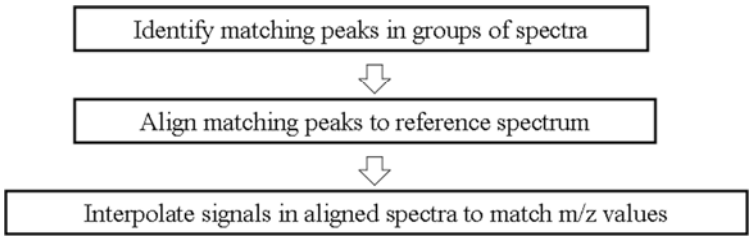


Figure 2.14 Steps involved in the alignment strategy developed in-house for PBS-IIc data.

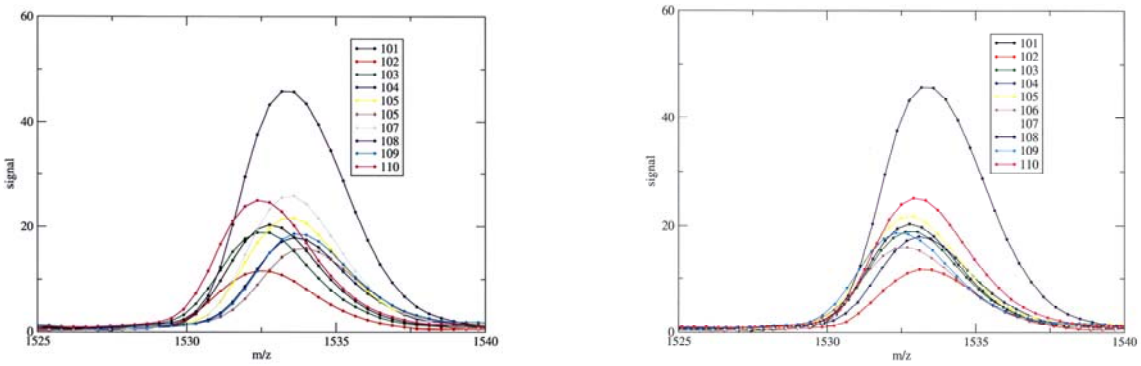


Figure 2.15 A peak 1533 depicted before (A) and after (B) alignment.

m/z	PBS-IIc (raw)		PBS-IIc (aligned)		prOTOF (raw)	
	Da	ppm	Da	ppm	Da	ppm
1011.06	1.47	1454	0.38	376	0.010	9.89
2255.12	2.34	1038	0.54	239	0.012	5.32
3156.64	2.97	941	0.65	206	0.024	7.60
4127.11	3.89	943	1.17	283	0.027	6.54
5903.79	4.25	720	0.80	136	0.034	5.76
9288.95	5.99	645	0.81	87	0.061	6.57

Table 2.1 The mass accuracy of six peaks across 1,000 to 10,000 Da is tabulated here in Daltons and parts-per-million for the unaligned (raw) and aligned data from PBS-IIc and unaligned prOTOF data.

In addition to the lack of mass accuracy, the PBS-IIc instrument also suffers from low resolution (resolution ~750-1,000) in the 1,000 to 10,000 window compared to the prOTOF (resolution 15,000-20,000). Moreover, spectral resolution of the lower resolution instrumentation may not be able to separate specific ions that are close in *m/z* and can coalesce multiple specific discreet ions into a single broad peak, with shoulders. The use of a shoulder peak *m/z* value in the final disease model will be difficult to reproduce accurately and may fail as robust marker events

[31]. The CV for peak intensity had previously been reported for the PBS-IIc as 16-26% while the prOTOF showed a CV of 5-10% [27], in agreement with the prOTOF profiles generated from our ProteinChip-MS platform. This and the less than 10 ppm mass accuracy observed indicate that replicate spectra from individual samples are reproducible in our workflow.

In short, we have shown the desired higher mass accuracy and lower mass drift can be attained from the high-resolution prOTOF, but not the PBS-IIc, the original system developed for ProteinChip analysis. To our knowledge, this is the first comparison between the performance of the prOTOF and PBS-IIc systems reported from the same sample spots. Consequently, data analyses were performed only on spectra acquired on the prOTOF instrument.

2.4 MODULE III: DATA ANALYSIS

This section covers discussion on the issue of overfitting, a novel consensus model approach, the statistical methods of logistic regression, CART, UPGMA, and t-test, and diagnostic accuracy measures. Given new discoveries are dominated by “70% successful” biomarkers [34], the focus is now to achieve higher sensitivity and specificity when these independent markers are considered collectively. SELDI is the embodiment of this model through its quantitative readout of multiple analytes that are combined into mathematical classification models. Indeed, the development of statistical algorithms for selecting promising biomarkers from a large pool of biomarkers is an active area of research [35-37].

2.4.1 Overfitting as a bias

Even though the literature is bombarded with reports of molecular markers for an assortment of diseases, an overwhelming majority of them remain insufficiently validated for clinical application. This reflects the difficult problem of determining which biomarkers warrant the

substantial investment of time and money required for validation efforts. This is especially true of biomarkers that arise from SELDI studies, partly due to the combination of non-robust data analysis approaches to underpowered studies that result in false positive findings.

Proteomics studies face the curses of high dimensionality and sparse data [10], making it crucial to recognize the issue of overfitting in the search of biomarkers. The studies generate high dimensional data where the number of variables (mass peaks) far exceeds the number of independent samples analyzed, which no existing traditional statistical or computational tool can handle. In fact, discovery-based research clearly violates the rule for predictive models that dictate at least ten observations for each variable to bestow confidence in the results [38].

A common and frustrating occurrence in proteomic biomarker discovery is when different laboratories studying the same disease, and employing different statistical methods on the same data set, end up producing non-overlapping sets of biomarkers. To date, no guidelines exist that facilitate the selection of appropriate statistical methods to employ for data analysis in mass spectra. Consequently, the choice of a statistical platform for each study remains subjective. All data analysis methods have their strengths and weaknesses but the caveat lies in the realization of their statistical power only when applied to data sets where the underlying data distribution assumptions are met. In the case of mass spectrometry data, no *a priori* knowledge of data distribution is available. Consequently, various learning algorithms have been engaged in the field for classification purposes, each with its underlying biases and assumptions of distribution [35-37].

Validation on independent data sets, independent of those used for discovery, is necessary to avoid overfitting. However, this can prove to be challenging due to limited sample availability. Here we describe a novel data-mining approach for the analysis and interpretation of mass spectra data to uncover truly discriminatory biomarkers as a solution to the issue of overfitting.

2.4.2 Consensus approach

The methodology described here can be applied to any MALDI TOF derived data set for any disease, provided the same standard operating procedure (from biological sample procurement, processing, and complexity reduction to actual mass spectrometry data acquisition) is employed. In the proposed workflow, both parametric (logistic regression, hierarchical clustering, t-test) and non-parametric (Classification and Regression Tree - CART) approaches were adopted to analyze the raw peaks from our data set to obtain a set of consensus biomarkers (Fig. 2.16). These four statistical platforms were selected because they were available either as licensed softwares that accompanied the mass spectrometers or were developed in-house. In parametric approaches, the data are assumed to originate from variables with a certain probability distribution (such as normality and homoscedasticity). Non-parametric approaches are more robust and yield greater power with less well-behaved data since no prior assumptions are made.

Consensus biomarkers are loosely defined as mass peaks with discriminatory power between the groups being compared that end up on the list of statistically differential peaks across at least two or more of the statistical strategies employed in the data mining analysis. Ideally, the most discriminatory marker peaks are selected as differential by all the methods employed. The reasoning is that in lieu of the data distribution knowledge, mass peaks that survive stringent conditions across multiple statistical methods are more likely to be true “biomarkers” and not artifacts as a consequence of bias inherent to a particular algorithm. Convergence upon a distinct set of biomarkers using multiple analytical platforms will confer higher confidence in these markers as robust entities and will increase the chance these markers may be adopted as diagnostic entities where subsequent identification and validation efforts should be directed. These biomarkers are either present specifically (all or none) or preferentially (relatively higher in one of the groups). Candidate peaks should be divided equally among peaks that increase and decrease with disease to minimize effects of variation in absolute signal intensities.

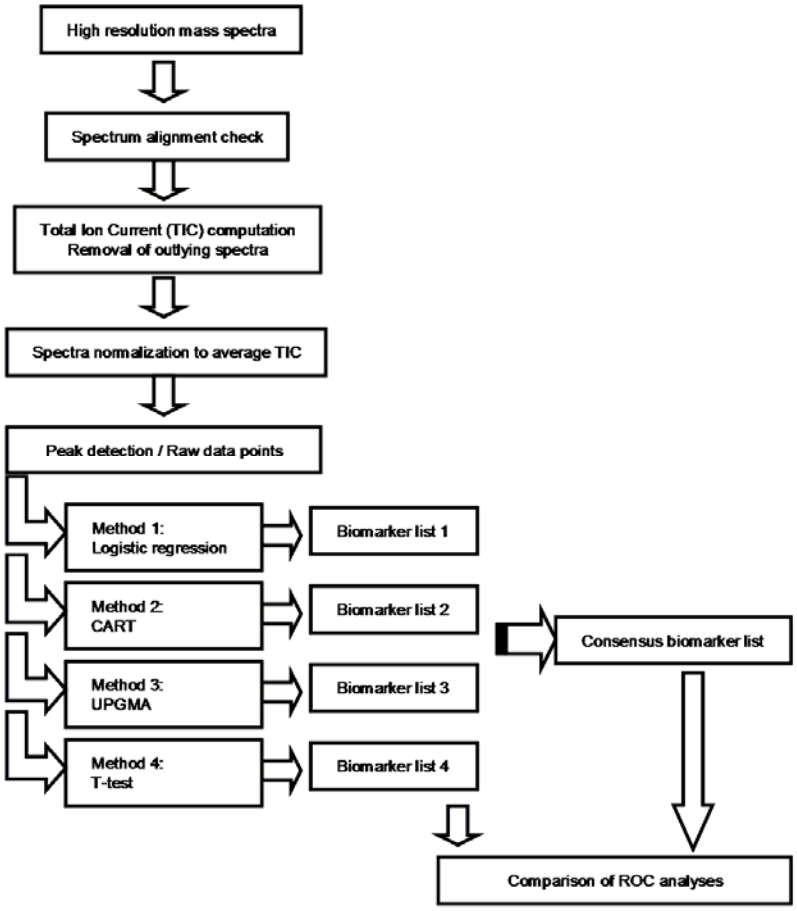


Figure 2.16 Schematic diagram of the multi-statistical workflow to discover consensus biomarker peaks.

This approach is novel in several respects. First, comparisons of different statistical methods on the same mass spectrometry data have been reported previously [39-41] but the ultimate goal of these reports was the selection of a method whose prediction model outperforms the rest of the methods under investigation when applied to a given set of experimental data and the subsequent recommendation of the method that prevailed for future analyses. This introduces bias in the selected marker peaks which are unique to a statistical method and are most often a result of overfitting. This is also true when peak reduction was performed using a predefined statistical method prior to submitting the remaining peaks for model building comparisons. In the

data analysis workflow described here there is no biased peak selection prior to model building by the four statistical algorithms under investigation; all raw peaks within the 1,000 to 10,000 mass range were subject to each algorithm.

Second, we used a mass spectrometer with high mass accuracy and low mass drift to generate high-resolution data, which is essential for accurate peak-to-peak comparison across spectra. A majority of the previous studies were performed using low-resolution mass spectrometer data with significant mass drifts across spectra within a single experimental run that further complicate analysis.

Finally, to assure a fair comparison between the methods during validation of this data analysis approach, the best discriminatory peaks from each method underwent the same diagnostic accuracy testing via receiver operating characteristic (ROC) curve analysis, as did the model consisting of only the consensus peaks. ROC confers a better sense of diagnostic performance of the biomarker peaks as it evaluates all possible cutoff values and produces the best trade-off between the rates of false-negative and false-positive results. The results from this validation stage of the consensus approach are reported later in this chapter when the consensus model is discussed (Section 2.4.9).

2.4.3 Data preprocessing

All the spectral files were first processed to restore the repeating zero signal values removed by the instrument software. During this process, the *m/z* lists were harmonized so that every spectrum has the same list of *m/z* values. In previous studies, we found that when comparing groups with relatively few spectra it was beneficial to smooth the raw data. In these instances, the Savitsky-Golay smoothing algorithm using a 9-point fit to a cubic function is applied. High frequency noise in the spectrum is reduced by the smoothing but fidelity to the major features is preserved.

An inherent challenge in analyzing mass spectral data is that they suffer from high dimensionality, and chemical and biological noise [10]. Therefore, in an attempt to reduce dimensionality, the *m/z* peak list was trimmed and only peaks that fall within the *m/z* range 1-10 kDa were subject to statistical analysis. We chose the lower cutoff value of 1 kDa to exclude any potential chemical noise contributed by the ionizing matrix. Furthermore, some proteins or peptides might be too small to be biologically informative [42]. The upper cutoff value of 10 kDa was selected because the ionization efficiency of molecules decreases with increasing mass and few peaks above the noise level were detected beyond this value.

The spectrum-to-spectrum alignment was checked for 6 different peaks across the *m/z* range of 1-10 kDa and found to be acceptable (<10 ppm). Due to the high mass accuracy and minimal mass drift of the prOTOF observed here and in agreement with others [27, 43, 44], no further spectral alignment was necessary.

The total ion current (TIC) of each spectrum is calculated and the average TIC was computed across all spectra. Spectra with a TIC value that was either twice or half of the average TIC were deemed outliers and were omitted from the study. Global normalization of the signal intensity of the mass peaks was performed by normalizing to the average TIC of the remaining spectra. This confers a sense of commonality across spectra for statistical comparisons. All spectra were run through the Progenesis PG600 software (Nonlinear Dynamics, UK) for peak detection using the following parameters to remove background noise: noise filter size 4, background filter size 70, and isotope detection in MALDI mode with peak threshold 25 and window 0.1 Da.

2.4.4 Logistic regression

Logistic regression is a parametric modeling technique that can be used to estimate the probability that an individual would acquire a complex disease [45]. It produces the most

parsimonious model (incorporating the minimum number of variables necessary) to explain the observations or to categorize the disease and control groups. Logistic regression does not require a normal distribution and homoscedasticity for the outcome variable. Instead, it assumes the outcome has a binomial distribution and is governed by the logistic function. Since proteomic data provide a large number of variables relative to the number of observations resulting in the problem of sparse data mentioned earlier, inaccurate estimates of the parameters needed to predict the status of the new subjects can result. To address this issue, a more detailed 9-step protocol was developed based on recommendations by the Environmental Protection Agency and SAS User Group International publications to replace the default one-step calling of the PROC LOGISTIC procedure in SAS.

All the variables (*m/z* values representative of peaks) from the data set were run first through a univariate analysis to test for significance in predicting the outcome of the samples. These variables are then checked for correlation. Since logistic regression assumes no collinearity among its variables, each pair of correlated variables will have the less significant one removed based on the univariate analysis. This trimming of the number of variables per observation is necessary to reduce dimensionality. In addition, variables with a Variation Inflation Factor exceeding 10, indicating multicollinearity were also removed [46].

Modeling was performed using the stepwise procedure, where variables were added and/or removed at each step depending on a significance test or some measure of information contributed by that variable to the difference between the groups. Here we set the significance level of entering, SLENTRY, to 0.990 and for staying, SLSTAY, to 0.995. This procedure continued until no variables can be added or removed. The stepwise technique effectively reduces the number of models under consideration while the less stringent entry and stay criteria allow more variables to be considered concurrently. Although stepwise procedures rely on tests or information for a single variable, all decisions are based on multivariate analyses. The model with the lowest Akaike Information Criterion (AIC) score will indicate the optimal number of

variables (n) to be incorporated in the model. AIC is a fitness function used to score models based on their quality to describe a data set. The best regression model is the one that minimizes the criteria used. AIC was preferred over the Schwarz Information Criterion (SC) that also accompanies logistic regression in SAS because it is better suited for the current goal of prediction [47]. Subsequent modeling will then produce a list of potential models incorporating n-2, n-1, n, n+1, and n+2 variables. The best subset selection method was coupled to the AIC analysis to incorporate suboptimal models that flank the optimal model with the lowest AIC. Only models with high Hosmer-Lemeshow (Goodness of Fit) score and low AIC score will be retained. They then undergo diagnostic checking to identify outlier observations and interaction between variables. A typical output from logistic regression is shown in Figure 2.17.

Logit(π) = 39.2625 - 0.0302(M3986_99) - 0.0855(M2225_14) - 0.1119(M1431_80) + 0.0643(M1839_98) - 0.0185(M5857_74)

Response Profile		
Ordered Value	GROUP	Total Frequency
1	1	19
2	0	45

Model Fit Statistics		
Criterion	Intercept Only	Intercept and Covariates
AIC	79.849	**57.798**
SC	82.008	70.751
-2 Log L	77.849	45.798

Hosmer and Lemeshow Goodness-of-Fit Test		
Chi-Square	DF	Pr > ChiSq
4.4196	9	**0.8817**

Odds Ratio Estimates			
Effect	Point Estimate	95% Wald Confidence Limits	
M3986_99	0.970	0.944	0.998
M2225_14	0.918	0.858	0.982
M1431_80	0.894	0.836	0.956
M1839_98	1.066	1.010	1.126
M5857_74	0.982	0.970	0.993

Association of Predicted Probabilities and Observed Responses			
Percent Concordant	91.0	Somers' D	0.820
Percent Discordant	9.0	Gamma	0.820
Percent Tied	0.0	Tau-a	0.348
Pairs	855	c (ROC)	**0.910**

Table of GROUP by pred_dis			
GROUP	pred_dis		
Frequency Percent Row Pct Col Pct	0	1	Total
0	43 67.19 ***95.56*** 84.31	2 3.13 4.44 15.38	45 70.31
1 Spec NPV	8 12.50 42.11 15.69	11 17.19 ***57.89*** 84.62	19 29.69
Total	51 79.69	13 20.31	64 100.00

Percent Correctly Predicted (for cutoff of 0.62) = 84.38

Figure 2.17 Model parameters from logistic regression.

We used our modified, AIC-optimal logistic regression protocol to analyze the data set generated in our narcolepsy case study (Chapter 4) and compared the diagnostic power of the best model from this approach to the best model obtained using the default single-step calling of the

PROC LOGISTIC in SAS. The diagnostic measures are shown in Table 2.2. The final model from the default stepwise procedure has a higher AIC statistic of 69.646 with two variables incorporated while the AIC-optimal model from the modified procedure has an AIC statistic of 57.798 with five variables incorporated. This means that the default stepwise model incorporated three predictors less than necessary to form a better predictive model. This is indicated by its lower Hosmer-Lemeshow goodness of fit statistic (0.669 for default versus 0.882 for AIC-optimal). The resulting default model also has a poorer discriminatory power than the AIC-optimal model as indicated by the lower area under the ROC curve. As for diagnostic accuracy (covered in more detail later in this chapter), both models have comparable sensitivity, positive predictive value (PPV) and negative predictive value (NPV), but the default model lacks in specificity and the percentage of cases accurately predicted (Table 2.2). This demonstrates that the modified procedure performs better than the default in producing good predictive models, and was thus adopted in subsequent logistic regression analyses. ROC curve analysis is performed on the few surviving models and an optimal cutoff value that gives the best sensitivity, specificity, and prediction accuracy is determined. Peaks from these models collectively create a pool of potential biomarker candidates.

Logistic Regression Model	Default	AIC-Optimal
Number of variables in final model	2	5
Goodness of Fit	0.669	0.882
AIC Statistic	69.65	57.80
Area under ROC curve	0.793	0.910
Sensitivity (%)	63.16	57.89
Specificity (%)	82.22	95.56
PPV (%)	85.96	84.62
NPV (%)	84.09	84.31
Percent accuracy (%)	76.56	84.38

Table 2.2 **Diagnostic accuracy measures from the default and AIC-optimal models in logistic regression**. AIC= Akaike Information Criterion, ROC= Receiver Operating Characteristic, PPV = Positive Predictive Value, NPV= Negative Predictive Value.

An advantage of modeling using logistic regression is its ability to estimate the associated risk with each variable. A drawback is that logistic regression suffers from the inability to accurately estimate the needed parameters when the two groups are perfectly separated based on the variables included in the model. Small sample sizes will also render the estimates unstable. This situation, however, will not be encountered in CART. Logistic regression was performed using the Statistical Analysis Software (SAS) (SAS Institute Inc., Cary, NC).

2.4.5 Classification and regression tree (CART)

CART is another analytical method that can be used to generate a prediction model. It is non-parametric, non-algebraic, and is a form of binary recursive partitioning where each group of patients at each "node" in a decision tree can only be split into two groups [48]. The construction of the classification tree begins with the variable that maximizes the group homogeneity of the daughter nodes. This process is then repeated where every daughter node is split into two subgroups until all variables have been exhausted or the end nodes are homogeneous (Fig. 2.18). The process involves the estimation of several linear combinations of predictor variables by discriminant function analysis of computing classification scores (or probabilities) that allow for prediction or classification of cases. Variable selection at each node is performed with one of six criteria – Gini, Symgini, Twoing, Ordered Twoing, Class Probability, or Entropy using Ciphergen's Biomarker Patterns Software (BPS).

Since it is non-parametric, no assumptions are made about the underlying distribution of the variables and highly-skewed, non-normal data sets can be handled. A drawback is some models can be unstable. Since all possibilities are evaluated at each splitting node, there is the potential of overfitting the model. To account for this, the tree is then pruned back using 10-fold cross-validation to obtain the optimal tree with the lowest average decision cost or error rate (Fig. 2.19).

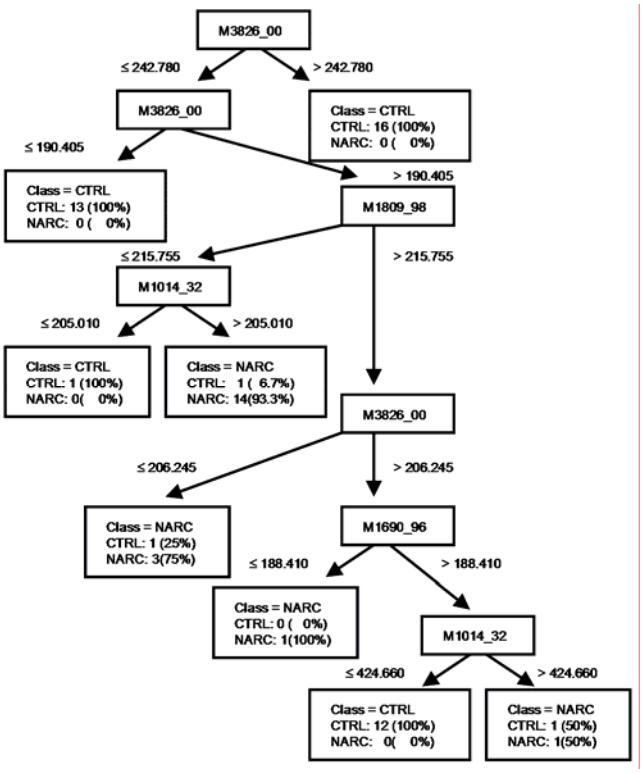


Figure 2.18 Tree diagram from CART analysis.

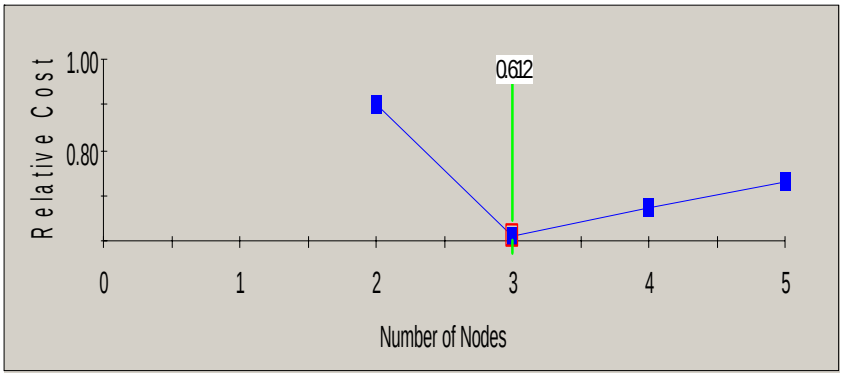


Figure 2.19 **Selection of best model from CART analysis**. Of the four models shown here, the best model with three nodes has the lowest error rate or cost.

2.4.6 Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

UPGMA is a hierachical clustering algorithm based on the average dissimilarity, or distance, between the clusters and their correlation. UPGMA is a parametric technique that is most commonly used in microarray [49, 50] and mass spectrometry data analysis [35] because no prespecification of number of clusters is required. Each group is compared and a p-value is automatically calculated for each peak using ANOVA based on the spectra groups. The resulting discriminant markers between the two groups depend on the stringency parameters for biomarker selection, such as minimum peak intensity and p-value (Fig. 2.20). This ensures only peaks that are above the noise level and are statistically significant are selected. UPGMA clustering was performed on the Progenesis PG600 software (NonLinear Dynamics, UK).

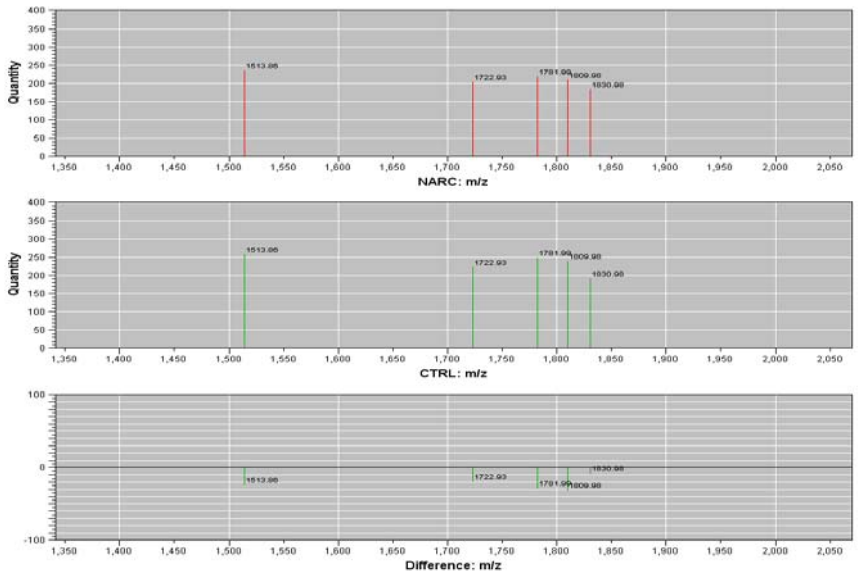


Figure 2.20 **Differential display of candidate marker peaks in the two sample groups in UPGMA**. Top two panels show the average signal intensity of the marker peaks in the respective sample group. Bottom panel shows the magnitude of the difference for each peak.

2.4.7 T-test

A high-throughput software pipeline developed in the Garner Lab at UT Southwestern was also used to analyze the data sets. This analysis is similar to a previously published method [44] but uses a t-test instead of the Cohen's d statistic used in the published method. The method using either the t-test or d statistic yielded very similar results. This software is a parametric, non-algebraic method for finding differential marker peaks by applying three filters to the average intensity value of each raw *m/z* data point between the two groups being compared. The first criterion uses a t-test for measuring the difference between the means. Typically, two signals whose means differ with a p-value of 0.05 or less are deemed significantly different. The second criterion requires the signals to be above the noise level. The third criterion requires the ratio between the two signals to be above a preset threshold (a so-called fold change determinant). Signals that pass all three filters suggest that a difference exists between the two groups being compared (Fig. 2.21).

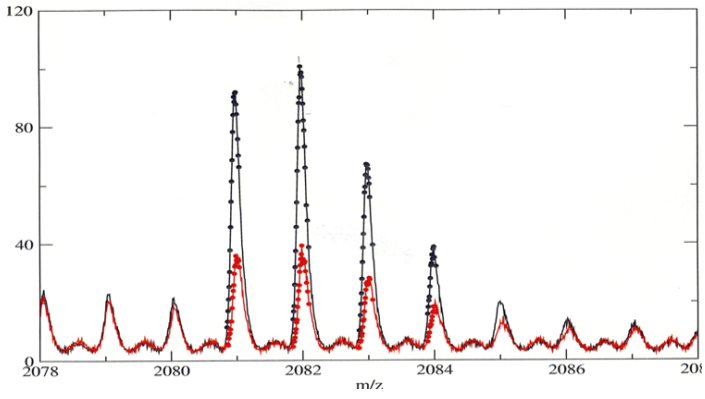


Figure 2.21 **A differential peak found in the in-house T-test method**. The two groups compared are represented by different colors. The circles on the peaks indicate differential data points that are statistically significant.

An advantage of this approach is that it does not require peak finding and thus is applicable to spectra with overlapping or non-Gaussian peaks, conditions that would confound most peak finding algorithms. Further, the method automatically provides a weighting factor for each peak, as peaks that differentiate the most have more discriminating data points on them [44].

2.4.8 Diagnostic accuracy measures

All statistically differential biomarkers discovered must be evaluated to determine their discriminatory performance characteristics between disease and non-disease patients as p-value alone does not indicate clinical utility. Diagnostic performance can be evaluated from its accuracy and predictability [51].

Diagnostic accuracy establishes how accurately the test discriminates between those with and without the disease and can be determined by calculating its sensitivity, specificity, and ROC curve [51]. Sensitivity is a measure of the ability of the test to identify a condition when it is present (Table 2.3). A high sensitivity corresponds to a low false negative rate (Type II error). False negatives are concerning as they could lead to the misled diagnosis of disease and missed opportunity for therapeutic intervention. Specificity is the ability of the test to rule out a condition when it is absent. A high specificity corresponds to a low false positive rate (Type I error). Maintaining high specificity (low false-positive rates) is a very high priority as even a small false positive rate translates into a large number of people subject to unnecessary costly diagnostic procedures and unwarranted distress. Although not always true, improving the sensitivity of a test may lead to decreasing its specificity, particularly if it involves choosing the threshold value for calling the test positive. Sensitivity and specificity are inherent properties of the biomarker test, and if well established, they will hold true regardless of the population tested. Depending on the clinical applications (e.g. screening or confirmatory diagnostics), different diagnostic measures

will take precedence. For example, high sensitivity is needed for screening but high specificity is needed for confirmation and subcategorization. Ideally, both need to be greater than 70% [34].

Test	Disease		TOTAL
	Present	Absent	
Positive	A = true positives	B = false positives	A+B = Test positives
Negative	C = false negatives	D = true negatives	C+D = Test negatives
TOTAL	A+C = Diseased	B+D = Nondiseased	A+B+C+D= Total samples

Table 2.3 **Contingency table for diagnostic accuracy measures**. Sensitivity = true positive rate = A/(A+C), Specificity = true negative rate = D/(B+D), PPV = A/(A+B), NPV = D/(C+D), Disease prevalence = (A+C)/(A+B+C+D).

ROC curves are derived from the calculated sensitivity and specificity values. The ROC curve, a statistical analysis method developed in the 1950s for evaluating radar signal detection, is used routinely today to evaluate diagnostic tests and generally for evaluating the accuracy of a statistical model (predictive models). Mean value comparisons between disease and normal may not be representative of the range of anticipated values in the population, which need to be separate enough that sample assignment is unambiguous so that an ideal cutoff value can be assigned. Given that there is no perfect test, there is an overlap in the value measured between the two groups, resulting in false positive and false negative results (Fig. 2.22). ROC analysis on all samples assists in the selection of the optimal cutoff value that represents a compromise between the total number of positive and negative results that can then be validated in separate data set. Advantages of ROC curve over simple frequencies and summary statistics for raw biomarker data are (i) it does not depend on the scale of raw-data measurements, which greatly facilitates comparison of the discriminatory capacities of different markers and (ii) that it displays true- and false-positive rates, quantities that are more relevant for screening purposes than raw biomarker value themselves [52].

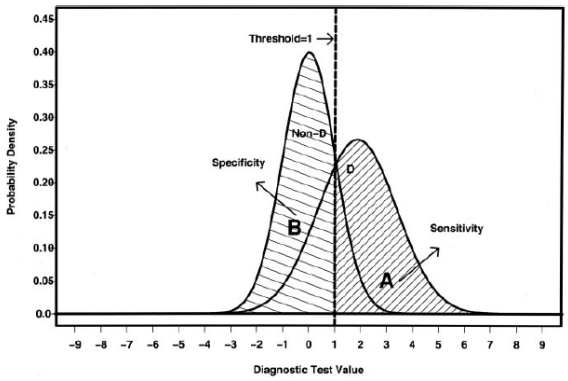


Figure 2.22 **Sensitivity, specificity, and cutoff value.** The specificity of the diagnostic test is represented as the shaded area under the non-disease distribution (A) above the arbitrary cutoff value (threshold). Sensitivity is represented as the shaded area under the disease distribution (B) below the same cutoff value. Both the sensitivity and specificity vary accordingly, with lower sensitivity and higher specificity as the threshold increases. [53]

Graphically, a ROC curve is a plot of sensitivity on the y-axis against (1-specificity) on the x-axis for varying values of the cutoff value (Fig. 2.23). Area under the ROC curve (AUC) is frequently used to describe a test's validity. The AUC is an overall summary of the diagnostic accuracy across the spectrum of the test. An AUC of 1.0 means the test has almost perfect discriminatory power between the groups compared for the diagnosis of interest (line connecting (0,0) to (0,1) and (0,1) to (1,1) in Figure 2.23. As a rule of thumb, an AUC >0.7 is considered good while >0.8 is considered great discriminatory power. Only satisfactory discrimination should warrant the evaluation of diagnostic accuracy measures such as sensitivity, specificity, PPV and NPV. The 45° diagonal line connecting (0,0) to (1,1) is the ROC curve corresponding AUC=0.5 (random chance). A diagnosis based on a test with an AUC of <0.5 is deemed not useful and efforts toward its development should be terminated. In general, ROC curve analysis helps select optimal tests and discards suboptimal tests.

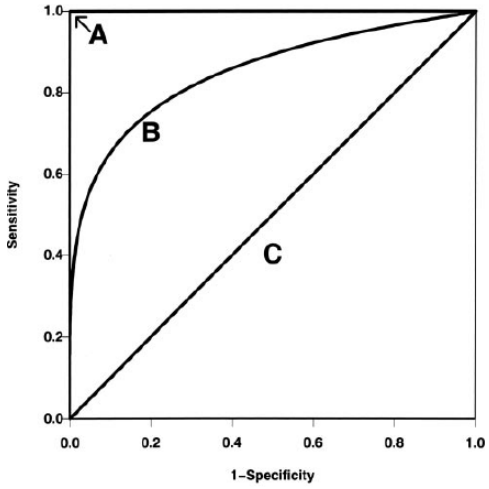


Figure 2.23 **ROC curve.** Three hypothetical ROC curves representing the diagnostic accuracy of the gold standard (A, AUC = 1), a typical ROC curve (B, AUC = 0.85), and a diagonal line corresponding to random chance (C, AUC = 0.5). As diagnostic test accuracy improves, the ROC curve moves toward A, and the AUC approaches 1. [53]

Diagnostic predictability establishes the ability of the test to predict the presence or absence of disease for a given test result and is determined by calculating the positive and negative predictive values [51]. PPV represents the likelihood that a patient actually has the disease. NPV is the likelihood that the patient is actually disease free. The PPV and NPV will vary depending on the prevalence of the disease in the population that is being tested. For example, PSA with a sensitivity of 70% and a specificity of 90% when applied to a population of 100,000 will have a PPV of 88% if the population is nodule-positive but only 0.2% for general screening. Some tests are only valuable when applied to a specific population. Therefore, biomarkers derived from proteomics studies should be applied to a population that has a larger representation of the sample set used in discovery in order to maintain its predictability.

2.4.9 Consensus model

As validation of the consensus model approach, the four algorithms described above were applied to our pilot narcolepsy study. Details pertaining to this study are documented in Chapter 4 (Case Study II: Narcolepsy). The performance of the discriminatory peaks in the resultant models from logistic regression and UPGMA hierarchical clustering was evaluated via ROC analyses using SAS. The diagnostic accuracy measures of interest are the sensitivity, specificity, PPV, NPV, prediction accuracy, and the AUC for narcolepsy classification. These parameters were obtained from the CART models via 10-fold cross-validation using the BPS software, and from the t-test models based on the distance proximity of the differential data points from the spectrum to be classified to those from spectra with known classification [44].

In the comparison between narcoleptic and non-narcoleptic samples, four optimal models were obtained using the AIC-optimal logistic regression procedure, as listed in Table 2.4. The mass peaks from these four models were pooled as potential biomarkers selected from logistic regression.

Logistic Regression Model	1	2	3	4	Pooled
Number of variables	5	1	2	2	9
Mass peaks (*m*/z)	1431.80	1809.98	1809.98	1722.93	1431.80
	1839.98		3826.00	1740.94	1722.93
	2225.14				1740.94
	3986.99				1809.98
	5857.74				1839.98
					2225.14
					3826.00
					3986.99
					5857.74

Table 2.4 Discriminatory mass peaks from AIC-optimal models in logistic regression analysis on narcolepsy data set.

The best tree from CART analysis is the tree with the lowest cost across all splitting criteria. This optimal tree with a cost of 0.322 was obtained from the Twoing criterion and the diagnostic measures of this model are listed in Table 2.5.

CART	Optimal Model
Number of variables in final model	6
Mass peaks (*m*/z)	1014.32, 1690.96, 1809.98, 3043.43, 3826.00, 3986.99
Area under ROC curve	0.984
Sensitivity (%)	78.95
Specificity (%)	88.89
PPV (%)	75.00
NPV (%)	90.91
Percent accuracy (%)	85.94

Table 2.5 Diagnostic accuracy measures of optimal CART model.

The maximum p-value was set to 0.05 and the minimum signal intensity ratio was set to 1.5 in the in house t-test analysis. Only three possible candidate biomarkers were identified in the spectra – 1740.94, 3598.07, and 5078.90. They were all higher in the narcolepsy samples. The diagnostic performance of this three-peak model is shown in Table 2.6.

T-test	Optimal Model
Number of variables in final model	3
Mass peaks (*m*/z)	1740.94, 3598.07, 5078.90
Sensitivity (%)	33.30
Specificity (%)	84.20
PPV (%)	50.00
NPV (%)	72.70
Percent accuracy (%)	67.90

Table 2.6 Diagnostic accuracy measures of optimal t-test model.

Differential peaks from UPGMA clustering were selected with a p-value <0.05 and a fold-change of at least 10% between the two conditions being compared. A total of 3 peaks were obtained (Table 2.7) and their diagnostic accuracy measures were determined (Table 2.8).

Mass peak (*m/z*)	Fold change	p-value
1781.99	1.13	0.046
1809.98	1.15	0.007
3826.00	1.13	0.017

Table 2.7 **Statistically differential peaks from UPGMA model**. Peaks are presented with their respective fold change and p-value.

UPGMA	Optimal Model
Number of variables in final model	3
Mass peaks (*m/z*)	1781.99, 1809.98, 3826.00
Area under ROC curve	0.788
Sensitivity (%)	36.84
Specificity (%)	95.56
PPV (%)	77.78
NPV (%)	78.18
Percent accuracy (%)	78.13

Table 2.8 Diagnostic accuracy measures of optimal UPGMA model.

Even though the ideal scenario is to have consensus peaks across all four platforms, the two peaks that were considered truly robust in this study were mass peaks at *m/z* 1809.98 and 3826.00 which were selected as statistically differential in three of the four approaches. They were grouped collectively to form an independent diagnostic model. When used for diagnosis, these two peaks have a sensitivity of 63.16%, a specificity of 82.22%, a PPV of 85.96%, a NPV of 84.09%, and a percentage of cases correctly classified of 76.56%. The area under the ROC curve was 0.79 (Table 2.9).

	Consensus Model
Number of variables in final model	2
Mass peaks (*m/z*)	1809.98, 3826.00
Area under ROC curve	0.793
Sensitivity (%)	63.16
Specificity (%)	82.22
PPV (%)	85.96
NPV (%)	84.09
Percent accuracy (%)	76.56

Table 2.9 **Diagnostic accuracy measures of consensus model**. Consensus peaks included in this model are peaks selected as statistically differential across three of the four algorithms.

The diagnostic performance of the consensus peaks were evaluated collectively against the best individual model from each statistical method. In this study, UPGMA and the t-test produced predictive models that did not perform as well as those from logistic regression and CART, even though the model from UPGMA included the two consensus peaks. In contrast, the consensus model has the diagnostic potency in some diagnostic measures that is comparable to, if not better than, the individual models from each statistical platform (Fig. 2.24).

Although admittedly limited in sample size, our comparison between narcolepsy samples versus all non-narcolepsy samples in this pilot study served to emulate general population screening, which is the intended application of these biomarkers. In this case, sensitivity is of more importance than specificity. Logistic regression suffers from low sensitivity (57.89%). CART prevailed in these measures with 78.95% sensitivity and 88.89% specificity, followed by the consensus model with a reasonable sensitivity of 63.16% and a specificity of 82.22%. This is very encouraging as the current genetic marker for narcolepsy in general is based on the presence of HLA DQB1*0602 which itself only has a specificity of 40% [54]. Genetic markers confer susceptibility but are not ideal disease biomarkers as most people who are positive for the HLA DQB1*0602 gene do not develop narcolepsy. The more important diagnostic measures to

consider are PPV and NPV, which evaluate the applicability of the diagnostic test on the target population. The consensus model displays the highest PPV of 85.96% and an NPV of 84.09%, comparable to the best logistic model. CART has the highest NPV of 90.91% but lacks in PPV with only 75% (Fig. 2.24).

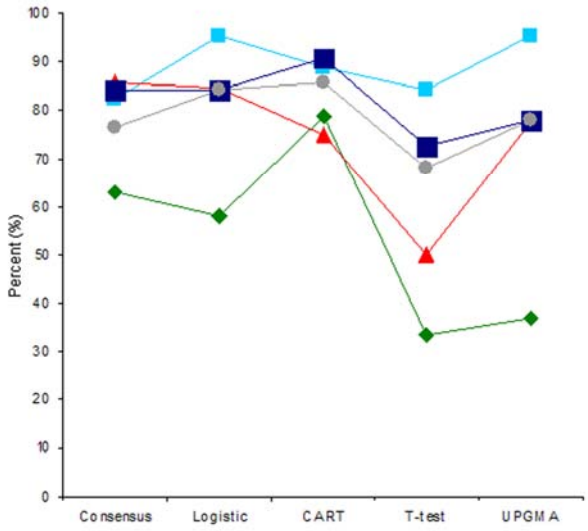


Figure 2.24 **Diagnostic measures comparison of consensus model to the best model from each of the four statistical approaches**. Green diamonds= sensitivity, light blue squares= specificity, red triangles= positive predictive value, dark blue squares= negative predictive value, gray circles= percent accuracy.

To summarize, CART seems to produce the best model in this pilot study when all five diagnostic accuracy measures are considered collectively, followed by logistic regression. UPGMA and the t-test did not fair as well. Of interest is the performance of the consensus model which seems to be a good compromise between both CART and logistic regression. Albeit models from logistic regression and CART in this study performed better in a few of the diagnostic measures, not all peaks in those models warrant subsequent identification and validation efforts. This is because spurious peaks that are only specific to those models might be a

reflection of overfitting and biases to the respective algorithm. These biases could be the reason why the logistic and CART models appear to perform better than the consensus model. Therefore, downstream validation efforts and resources will be better off directed at the consensus peaks.

An added advantage of the consensus model is the higher level of confidence in the true discriminatory traits of the peaks as they managed to survive various data distribution assumptions across statistical platforms to appear as statistically significant differentially expressed peaks. Another advantage of forming a consensus model is the trimming of the long list of potential biomarkers to be sequenced to the selected few with true discriminatory power. The ideal clinical assay will only need to focus on assaying the minimal number of biomarkers to accurately diagnose a disease state.

The methodology described here can be applied to any MALDI TOF derived data set to reconcile the disparate potential biomarker mass peaks reported by different studies on the same disease, provided the same standard operating procedure is employed during data acquisition. Consensus peaks will no doubt expedite efforts to identify robust biomarkers for clinical applications as their true discriminatory trait is reflected in their selection as differential biomarkers across several statistical platforms. Hence, they should be the main candidates where downstream identification and validation efforts should be focused on to assess their suitability to be adopted in a diagnostic assay or as therapeutic targets. Hopefully, the strategy proposed here will stimulate further advances and alternative approaches to the disease-state profiling based on high dimensional proteomic data and contribute to the discovery of useful biomarkers.

2.5 IDENTIFICATION AND VALIDATION STRATEGIES

Critics of the differential pattern profiling approach have argued that not knowing the structure of the biomarkers make validation more difficult because there is no physiological hypothesis to help give confidence in the findings [55]. This is unfounded because, from a historical context, the ovarian cancer marker CA-125 has been measured for years without knowing the underlying identity or amino acid sequence. Furthermore, PSA has been adopted as the marker for prostate cancer without knowing the underlying physiologic basis for this correlation. There does not necessarily have to be a causal link between markers and disease. What is necessary, however, is that the association of a particular molecular marker with the disease of interest be reproducible in a statistically robust manner within and between testing sites, and can affect disease management by influencing therapeutic options.

As efforts to ensure reproducibility are still ongoing, the immediate future calls for the identification of these putative markers to facilitate transition to the more established, higher-throughput assay platform of ELISA which requires antibody generation. This will allow for the immediate bias and overfitting assessment in clinical samples. Furthermore, candidate identity is important to promote biological insight into the molecular mechanism of diseases. This information will reveal their point of origin and physiology. It is only when the identity of the protein is known that new therapeutics can be generated to prevent the disease occurrence or to better the lives of patients who failed early detection. This represents a paradigm shift from an unbiased discovery approach emphasizing comprehensive protein characterization via MS to a candidate-driven approach emphasizing high-throughput quantitative antibody-based assays.

An advantage of our data analysis approach is it creates a sense of stringency in the discovery process by reducing the number of potential candidate biomarkers for downstream identification and validation efforts. The prioritization of the candidates for verification will generally be dependent on the individual marker’s performance.

A significant disadvantage of SELDI- and MALDI-based approaches is it allows for the relative quantitation of marker peaks but does not allow for their identification. Towards this end, various technologies with different strengths and weaknesses have been integrated to enrich for and identify the candidate markers.

2.5.1 Identification strategies

The identification of candidate marker peaks requires their enrichment and subsequent sequence determination. It is essential to enrich for the biomarker peaks to fully exploit the limited sensitivity of the mass spectrometers used for sequencing, especially if the peaks originate from low abundance proteins.

2.5.1.1 Biomarker enrichment

A popular strategy is to first recapitulate the chip chemistry onto columns. In our workflow, the IMAC chip will now be replaced by IMAC columns, preferably with the same ion chelating group used on the ProteinChips. The array chips have immobilized nitrilotriacetic acid (NTA) groups to chelate the nickel ions via a tetradentate metal capture strategy (Figure 2.25). Consequently, this leaves two free binding sites on nickel to capture proteins. Iminodiacetic acids (IDA) chelate metal ions through a tridentate capture configuration, allowing three binding sites on nickel to capture proteins. Even though more proteins can be captured by IDA, they undergo metal leaching due to weaker binding to the metal ion compared to NTA. These IMAC columns with suspended resins that enable overnight incubation for equilibrium binding to occur have a higher binding capacity and allows for more target proteins to be captured. The bound species can then be eluted with MS-compatible solvents such as imidazole (LC-MS/MS) or acetonitrile (MALDI-MS/MS). Identity-based biomarker discovery is reliant upon multidimensional fractionation at the protein and/or peptide level to improve detection of low abundance species.

Therefore, the recapitulated samples can then be used for biomarker enrichment into selected fractions using tandem orthogonal separation technologies, such as MudPIT. Coupling more than two sequential separation techniques will risk inefficient recovery of scarce biomarkers in the process. Micro- or nanoflow HPLC may be employed to concentrate the desired species in a smaller volume prior to MS analysis. Automated multidimentional separation system, such as the PF2D [56], can also be employed by pooling clinical samples that display strong marker peak signal intensity in the mass spectra. A potential hurdle here is that since our biomarkers are of LMW, the difference in hydrophobicity between the species might be too subtle to be fully resolved on a reverse phase column. This strategy also negatively impacts the throughput as even a small number of patient samples will produce large numbers of fractions for analysis. Gel separation of the complex samples prior to MS analysis is also a valid enrichment method if the marker peaks are not of LMW as gels are biased against species <10 kDa.

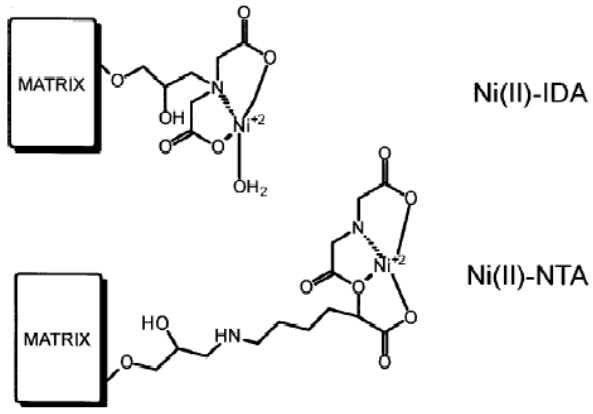


Figure 2.25 **Structures of nickel binding matrices**. Ni(II)-IDA allow three binding sites for protein capture and a tridentate metal capture configuration. Ni(II)-NTA allows two binding sites for protein capture and a tetradentate metal capture configuration.

2.5.1.2 Biomarker sequencing

Assuming the biomarkers are successfully enriched, these fractionated samples can then be analyzed by single stage MS to determine which ones contain the marker peaks. MS/MS sequencing can then be used to sequence these peaks for protein identification via database search by tandem MS sequencing using the 4700 TOF/TOF Proteomics Analyzer (Applied Biosystems Inc., CA). This system was designed for high throughput with possible automation of tandem data acquisition followed by protein database search using their integrated Global Proteome Search Explorer software. This instrument is routinely used for peptide mass fingerprinting and tandem MS protein sequencing. It has a mass accuracy of 10 ppm and sensitivity down to the low picomole range.

It is not easy to obtain efficient fragmentation of large molecules (proteins and large peptides) that allow for sequence determination. Known fragmentation processes including CID, ECD, ETD and IRMPD are in general biased toward smaller ions. The limitation of tandem MS sequencing is the tendency of fragmentation to occur at the few preferred sites on the molecules, resulting in only a few dominating fragment peaks that are not sufficient to obtain complete sequence information (Figs. 2.26 and 2.27). Ion trap or FT-ICR can be useful in this instance because of increased sensitivity. An ion trap, in theory, supports up to MS^n fragmentations. Therefore, the couple dominant fragment peaks in MS^2 can still be selected for further fragmentation to obtain a detailed sequence. The alternative approach to obtain the identities of these proteins in highly complex mixtures is to convert them enzymatically (usually by trypsin digestion) into their peptide components. However, a serum sample with 10,000 different proteins with mass range 10 to 100 kDa, each producing 20 peptides, will result in 200,000 peaks in MALDI within the 1 to 5 kDa mass range. Hence, separation of peptides for sequence determination is a must. This again leads to the problem of sample amplification from parent input to daughter fractions.

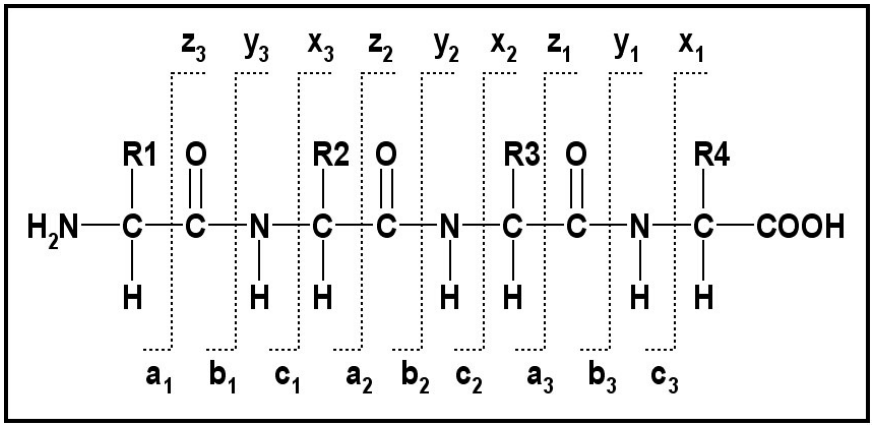


Figure 2.26 Collision induced dissociation of parent ions result in different sets of product ions, depending on the site of fragmentation.

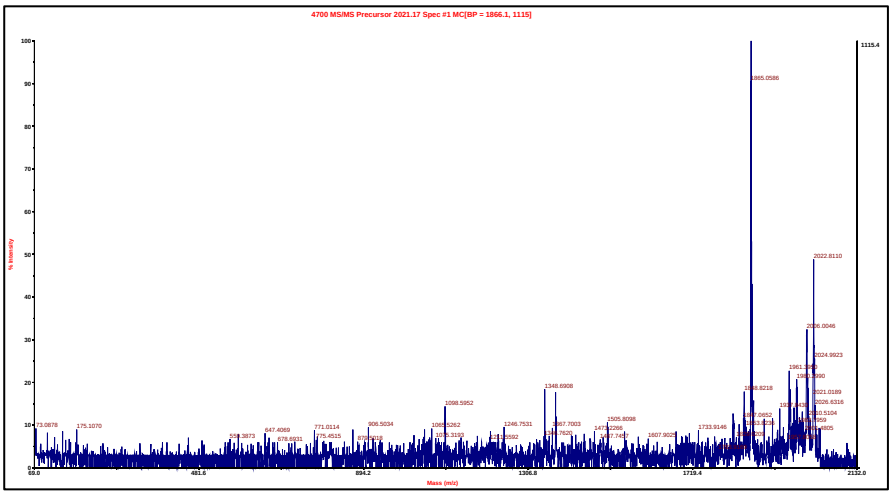


Figure 2.27 **Limitation of tandem MS analysis**. MS/MS analysis of a peak 2021 resulted in inefficient fragmentation that did not lead to sequence identification.

Another complementary approach is to subject the eluate from the IMAC columns to direct sequencing via FT-MS. This is possible through collaboration with sites that have access to the technology, such as ThermoFisher Scientific's BRIMS (Biomarker Research Initiatives in

Mass Spectrometry) Center in Cambridge, MA. This approach has proven successful in a recent report [57]. The beauty of this instrument lies in the coupling of an ion trap to the FT-ICR, both conferring superior sensitivity and resolution. FT-MS can reach the low attomole range in sensitivity. FT-MS has a unique resolution of 100,000 or more, and accuracies can be as good as 1 ppm. This confers the ability to perform specific searches on peptide mixtures obtained from small clinical samples.

Putative identity of the differential mass peaks can also be determined *in silico* by matching the nominal mass of the marker peaks to those in databases with sequence information. The confidence in the identifications is dependent on the mass accuracy of the mass spectrometer, with higher mass accuracies resulting in a more specific list of candidate proteins.

2.5.2 Verification and Validation strategies

A single study does not establish a scientific fact especially if the sample size is limited in size. Rather, secondary verification of results in the initial finding from the protein profiling study is imperative to show reproducibility and reduce false positives from overfitting. If the biological sample used in discovery is not the final diagnostic medium, verification in blood samples from the same patients facilitates transition to the eventual blood-based platform. Due to scarcity of samples, verification may have to occur in the same small sample set used in profiling. However, to assert confidence in the differential markers, a more targeted, and preferably orthogonal, assay will have to be used. The end result of this stage will be a reduced list of markers with high diagnostic performance that are now suitable for formal validation studies [51].

The validation stage will recruit hundreds of clinical samples, incorporating a broader range of cases and controls from different cohorts and geographic locations, to better capture the whole spectrum of variation in the tested population.

2.5.2.1 Antibody-based approaches

A common first step in verification is the generation of specific antibodies and their validation. In addition to using them for verification efforts, the antibodies can also be easily adapted to the sandwich enzyme-linked immunosorbent assay (ELISA) system, which is performed commonly in clinical laboratories. These tests are robust, high-throughput, and confer high sensitivity and specificity (pg/ml) because they use a pair of antibodies against the targeted molecule. However, use of an ELISA to test for the presence of a disease requires a single, meticulously validated protein biomarker of the disease, as well as extremely well-characterized, high-affinity antibody that can detect the protein of interest. Depending on the identity of the signature ion, it may or may not be feasible to proceed directly to develop a serum immunoassay for the individual biomarker. This is because the intensity of MALDI TOF MS does not reflect the concentration of the given marker associated with the ion. Moreover, cleaved versions and parent species might cross react with the antibody. A possibility exists to develop polyclonal antibodies as bait, and that following binding, the entirety of the recognized entities, including the diagnostic fragment are eluted and analyzed via MS. This is also advantageous to cover possible sequence changes in the same peptide in the heterogeneous human population notorious for single nucleotide polymorphisms (SNPs).

As an initial step, the biological samples can be subject to Western blot analysis using the antibody to confirm the modulation in the expression level of the antigens in question. It is also the first check for the specificity of the antibodies raised. Alternatively, the antibody can be used for immunoprecipitation experiments to deplete the biomarker in question. Western blot of the depleted and native samples will then confirm the expression level of the species and specificity of the antibody. A caveat though is that Western blots with serum require the tentative marker to be present in the µg/ml range. This may be overcome by pooling samples.

The beauty of SELDI is that this discovery platform can also be utilized to a limited extent for biomarker validation. One way to accomplish this is to immunodeplete the marker with

the antibody and run the depleted and non-depleted samples on the IMAC30 chips used in discovery to determine if the peak of interest vanishes in the depleted sample (Fig. 2.28). Alternatively, the antibody can be immobilized on a ProteinChip array, incubated with sample to capture the biomarker of interest, and then run on MS to see if the peak of interest is observed.

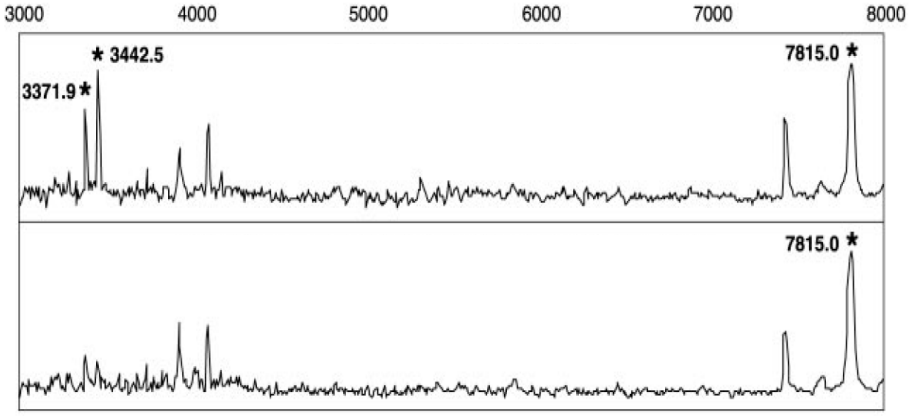


Figure 2.28 **Immuno-mass spectrometry approach for biomarker validation**. Immuno-depleted samples (bottom panel) when compared to their non-depleted counterparts (top panel) should witness an absence of the marker peaks (3371.9 and 3442.5 here) in MS analysis.

2.5.2.2 Antibody-free approaches

Candidate-based validation assays rely on the specificity of the capture or detection methods, as seen in ELISA. However, for most novel candidates, antibodies will not be available. The development of a reliable immunoassay for quantitation of one target protein is expensive, has a long development time, and is dependent upon the generation of high quality protein antibodies as mentioned. These reagent limitations along with the limited ability to multiplex immunoassays [58] make it necessary to use alternative targeted quantitative assays to bridge candidate discovery and validation.

Targeted, hypothesis-driven mass spectrometry using multiple reaction monitoring (MRM) is one such assay. MRM has long been a principal tool for quantification of small molecules in clinical chemistry [59-61]. MS-based quantitative assays is an attractive approach given their improved sensitivity and specificity, the speed at which assays can be developed compared to immunoassays and the quantitative nature of the assay with substantial multiplexing capability and precision (CV <5%) [62].

In MRM mode, the mass spectrometer with triple quadrupole capability transmits the parent ion and subsequent fragment ion with high sensitivity and selectivity. Figure 2.29 shows the configuration of the mass analyzers during MRM. The first mass analyzer is set to transmit only the mass of the peptide parent ion into the collision cell. Only one of the sequence ions of the peptide, generated by the fragmentation in the collision cell, is passed through the second mass analyzer to the detector. The detection of the peptide by MRM drives the acquisition of MS/MS to confirm the peptide sequence and thus definitively the identity of the detected peptide. From the theoretical peptide sequence, the fragmentation pattern of the peptide by MS/MS is predicted. When combined with chromatography, this makes the mass spectrometer a highly specific detector for the target molecule as it is highly unlikely that isobaric compounds that may coelute with the target compound will also have an identical fragment mass.

MRM coupled with stable isotope dilution MS (MRM/SID-MS) enables quantitation. Isotopically labeled standards are added to a sample in known quantities, and the signals from the exogenous labeled and endogenous unlabeled species are compared. The labeled molecules (either labeled enzymatically or synthesized to incorporate labeled amino acids) behave nearly identically to the unlabeled forms with respect to ionization efficiency. This approach enables a moderate number of candidate proteins (30–100) to be targeted simultaneously and measured in a statistically viable number of patient samples required for verification [63].

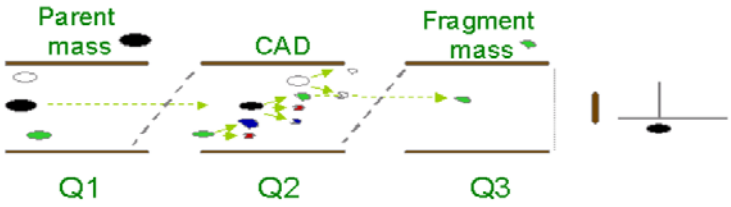


Figure 2.29 **Multiple Reaction Monitoring (MRM)**. Mass analyzer Q1 is set to only transmit the parent *m/z*. The collision energy is optimized to produce a diagnostic charged fragment of this peptide in Q2, and Q3 is set to transmit this diagnostic fragment only. Only precursor ions with this exact transition will be detected (Applied Biosystems).

Stable isotope standards and capture by anti-peptide antibodies (SISCAPA) [64] allows for the detection of biomarker candidates of low abundance by extending the sensitivity of the peptide assay by two orders of magnitude. In SISCAPA (Fig. 2.30), anti-peptide antibodies are used to enrich for the signature peptides before MRM analysis. Immunoaffinity peptide enrichment enhances both sensitivity and specificity, facilitating throughput by permitting analysis in complex matrices with little or no fractionation. Because the antibodies bind both the labeled and unlabeled monitor peptides equally (difference is in ^{13}C isotope content only), quantitative information is preserved throughout capture and elution. An additional level of specificity is conferred by the fragmentation pattern of the affinity-captured peptides, allowing SISCAPA to act much like an ELISA with MS/MS substituted for the second antibody that has absolute structural specificity. Although antibody production is required in this approach, the development time of single antibody enrichment approaches is typically faster than two-antibody ELISA assays [65, 66].

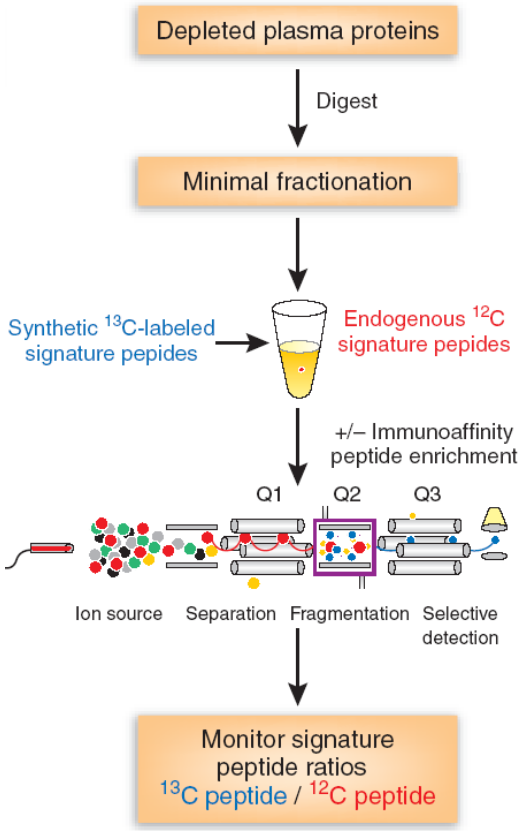


Figure 2.30 **LC-MRM/SISCAPA-MS workflow**. Depleted samples with minimal strong cation exchange fractionation and anti-peptide antibody enrichment significantly increases the MS-based assay sensitivity.

LC-MRM permits the ability to quantify over a wide dynamic range (over 5 orders of magnitude) and very low levels of detection (multiplexed assays for 6 proteins in plasma in the 1-10ng/ml range with CV from 3 to 15% without immunoaffinity enrichment of either proteins or peptides). Up to 1000-fold improvement compared with direct analysis of proteins in plasma by MS can be achieved by simple abundant protein depletion and minimal fractionation by strong cation exchange at the peptide level [8, 63, 67-70].

Priority is generally given to peptides detected in unbiased discovery experiments but not all peptides identified from a profiling study will be suitable for quantitative purposes due overlapping retention time and low confidence in peptide identification due to poor MS/MS

fragmentation. Therefore, in addition to the profiling markers, tryptic peptides can also be selected from *in silico* analysis of the protein sequence [71] to obtain several MRM peptides for the same protein. The peptides with high chances of usable MRM fragments will have to (i) demonstrate MS detectability with high MS/MS spectral quality, (ii) exhibit good chromatographic peak shape on reverse phase chromatography with preference for moderately hydrophobic peptides and (iii) be unique to the protein of interest. Since the semiquantitative measurements of peptides in the initial profiling study rely solely on the parent ion mass, they are susceptible to interference by unrelated peptides of the same *m/z*. The MRM method uses product ion intensity and has higher sensitivity due to double filtering of parent and product ion mass. Therefore, reproduction of the quantitative difference by MRM may serve to confirm the differential presence of the candidate markers.

2.6 CASE STUDIES

The workflow presented here serves as an addendum to the many innovative experimental designs in unbiased proteomics biomarker discovery to address the challenges discussed in Chapter 1. We demonstrate the practicality of this workflow by performing exploratory studies in two autoimmune diseases in Chapter 3 (Case Study I: Multiple Sclerosis) and Chapter 4 (Case Study II: Narcolepsy). The confirmation of differential presence of the discovered markers in an orthogonal platform for the respective study represents the fulfillment of our second objective of biomarker identification and verification. Our data-driven exploratory studies aim to identify proteins that are differentially present in confirmed disease relative to control samples, and prioritize identified markers in a reliable and reproducible manner. We further extended on this by estimating the diagnostic accuracy of the biomarkers via ROC curve analysis. This serves as a filter to direct only optimal markers for elaborate identification and verification efforts. Our

studies constitute the premise for further validation with a larger cohort derived from the target population across geographical sites.

2.7 BIBLIOGRAPHY

1. Petricoin EF, Ardekani AM, Hitt BA, Levine PJ, Fusaro VA, Steinberg SM, Mills GB, Simone C, Fishman DA, Kohn EC *et al*: **Use of proteomic patterns in serum to identify ovarian cancer**. *Lancet* 2002, **359**(9306):572-577.
2. Conrads TP, Fusaro VA, Ross S, Johann D, Rajapakse V, Hitt BA, Steinberg SM, Kohn EC, Fishman DA, Whitely G *et al*: **High-resolution serum proteomic features for ovarian cancer detection**. *Endocrine-related cancer* 2004, **11**(2):163-178.
3. Aivado M, Spentzos D, Alterovitz G, Otu HH, Grall F, Giagounidis AA, Wells M, Cho JY, Germing U, Czibere A *et al*: **Optimization and evaluation of surface-enhanced laser desorption/ionization time-of-flight mass spectrometry (SELDI-TOF MS) with reversed-phase protein arrays for protein profiling**. *Clin Chem Lab Med* 2005, **43**(2):133-140.
4. Sanders ME, Dias EC, Xu BJ, Mobley JA, Billheimer D, Roder H, Grigorieva J, Dowsett M, Arteaga CL, Caprioli RM: **Differentiating proteomic biomarkers in breast cancer by laser capture microdissection and MALDI MS**. *Journal of proteome research* 2008, **7**(4):1500-1507.
5. Han MH, Hwang SI, Roy DB, Lundgren DH, Price JV, Ousman SS, Fernald GH, Gerlitz B, Robinson WH, Baranzini SE *et al*: **Proteomic analysis of active multiple sclerosis lesions reveals therapeutic targets**. *Nature* 2008, **451**(7182):1076-1081.
6. WHO: **Use of anticoagulants in diagnostic laboratory investigations**. In.; 2002: 1-62.
7. Omenn GS, States DJ, Adamski M, Blackwell TW, Menon R, Hermjakob H, Apweiler R, Haab BB, Simpson RJ, Eddes JS *et al*: **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 3020 proteins and a publicly-available database. *Proteomics* 2005, **5**(13):3226-3245.

8. Han B, Higgs RE: **Proteomics: from hypothesis to quantitative assay on a single platform. Guidelines for developing MRM assays using ion trap mass spectrometers**. *Briefings in functional genomics & proteomics* 2008.

9. Sedlaczek P, Frydecka I, Gabrys M, Van Dalen A, Einarsson R, Harlozinska A: **Comparative analysis of CA125, tissue polypeptide specific antigen, and soluble interleukin-2 receptor alpha levels in sera, cyst, and ascitic fluids from patients with ovarian carcinoma**. *Cancer* 2002, **95**(9):1886-1893.

10. Somorjai RL, Dolenko B, Baumgartner R: **Class prediction and discovery using gene microarray and proteomics mass spectroscopy data: curses, caveats, cautions**. *Bioinformatics (Oxford, England)* 2003, **19**(12):1484-1491.

11. Zolg W: **The proteomic search for diagnostic biomarkers: lost in translation?** *Mol Cell Proteomics* 2006, **5**(10):1720-1726.

12. Guerreiro N, Gomez-Mancilla B, Charmont S: **Optimization and evaluation of surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry for protein profiling of cerebrospinal fluid**. *Proteome science* 2006, **4**:7.

13. Pieper R, Su Q, Gatlin CL, Huang ST, Anderson NL, Steiner S: **Multi-component immunoaffinity subtraction chromatography: an innovative step towards a comprehensive survey of the human plasma proteome**. *Proteomics* 2003, **3**(4):422-432.

14. Liu T, Qian WJ, Mottaz HM, Gritsenko MA, Norbeck AD, Moore RJ, Purvine SO, Camp DG, 2nd, Smith RD: **Evaluation of multiprotein immunoaffinity subtraction for plasma proteomics and candidate biomarker discovery using mass spectrometry**. *Mol Cell Proteomics* 2006, **5**(11):2167-2174.

15. Brand J, Haslberger T, Zolg W, Pestlin G, Palme S: **Depletion efficiency and recovery of trace markers from a multiparameter immunodepletion column**. *Proteomics* 2006, **6**(11):3236-3242.

16. Hirabayashi J: **Lectin-based structural glycomics: glycoproteomics and glycan profiling**. *Glycoconjugate journal* 2004, **21**(1-2):35-40.

17. Granger J, Siddiqui J, Copeland S, Remick D: **Albumin depletion of human plasma also removes low abundance proteins including the cytokines**. *Proteomics* 2005, **5**(18):4713-4718.

18. Hortin GL: **The MALDI-TOF mass spectrometric view of the plasma proteome and peptidome**. *Clinical chemistry* 2006, **52**(7):1223-1237.

19. Ekblad L, Baldetorp B, Ferno M, Olsson H, Bratt C: **In-source decay causes artifacts in SELDI-TOF MS spectra**. *Journal of proteome research* 2007, **6**(4):1609-1614.

20. Keller BO, Li L: **Discerning matrix-cluster peaks in matrix-assisted laser desorption/ionization time-of-flight mass spectra of dilute peptide mixtures**. *Journal of the American Society for Mass Spectrometry* 2000, **11**(1):88-93.

21. Smirnov IP, Zhu X, Taylor T, Huang Y, Ross P, Papayanopoulos IA, Martin SA, Pappin DJ: **Suppression of alpha-cyano-4-hydroxycinnamic acid matrix clusters and reduction of chemical noise in MALDI-TOF mass spectrometry**. *Analytical chemistry* 2004, **76**(10):2958-2965.

22. Krutchinsky AN, Chait BT: **On the nature of the chemical noise in MALDI mass spectra**. *Journal of the American Society for Mass Spectrometry* 2002, **13**(2):129-134.

23. Baggerly KA, Morris JS, Coombes KR: **Reproducibility of SELDI-TOF protein patterns in serum: comparing datasets from different experiments**. *Bioinformatics (Oxford, England)* 2004, **20**(5):777-785.

24. Diamandis EP: **Mass spectrometry as a diagnostic and a cancer biomarker discovery tool: opportunities and potential limitations**. *Mol Cell Proteomics* 2004, **3**(4):367-378.

25. Sorace JM, Zhan M: **A data review and re-assessment of ovarian cancer serum proteomic profiling**. *BMC bioinformatics* 2003, **4**:24.

26. Hortin GL: **Can mass spectrometric protein profiling meet desired standards of clinical laboratory practice?** *Clinical chemistry* 2005, **51**(1):3-5.

27. Albrethsen J: **Reproducibility in protein profiling by MALDI-TOF mass spectrometry**. *Clinical chemistry* 2007, **53**(5):852-858.

28. Marshall E: **Getting the noise out of gene arrays**. *Science (New York, NY* 2004, **306**(5696):630-631.

29. Conrads TP, Zhou M, Petricoin EF, 3rd, Liotta L, Veenstra TD: **Cancer diagnosis using proteomic patterns**. *Expert Rev Mol Diagn* 2003, **3**(4):411-420.

30. Mischak H, Apweiler R, Banks RE, Conaway M, Coon J, Dominiczak A, Ehrich JHH, Fliser D, Girolami M, Hermjakob H *et al*: **Clinical proteomics: A need to define the field and to begin to set adequate standards**. *Proteomics Clinical Applications* 2007, **1**(2):148-156.

31. Semmes OJ, Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, Cazares LH, Chan DW, Grizzle WE, Izbicka E *et al*: **Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility**. *Clin Chem* 2005, **51**(1):102-112.

32. Loboda AV, Ackloo S, Chernushevich IV: **A high-performance matrix-assisted laser desorption/ionization orthogonal time-of-flight mass spectrometer with collisional cooling**. *Rapid communications in mass spectrometry* 2003, **17**(22):2508-2516.

33. Yasui Y, Pepe M, Thompson ML, Adam BL, Wright GL, Jr., Qu Y, Potter JD, Winget M, Thornquist M, Feng Z: **A data-analytic strategy for protein biomarker discovery: profiling of high-dimensional proteomic data for cancer detection**. *Biostatistics (Oxford, England)* 2003, **4**(3):449-463.

34. Drukier AK, Grigoriev I, Brown LR, Tomaszewski JE, Sainsbury R, Godovac-Zimmermann J: **Looking for Thom's biomarkers with proteomics**. *Journal of proteome research* 2006, **5**(8):2046-2048.

35. Meunier B, Dumas E, Piec I, Bechet D, Hebraud M, Hocquette JF: **Assessment of hierarchical clustering methodologies for proteomic data mining**. *Journal of proteome research* 2007, **6**(1):358-366.

36. Ball G, Mian S, Holding F, Allibone RO, Lowe J, Ali S, Li G, McCardle S, Ellis IO, Creaser C *et al*: **An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers**. *Bioinformatics (Oxford, England)* 2002, **18**(3):395-404.

37. Ressom HW, Varghese RS, Abdel-Hamid M, Eissa SA, Saha D, Goldman L, Petricoin EF, Conrads TP, Veenstra TD, Loffredo CA *et al*: **Analysis of mass spectral serum profiles for biomarker selection**. *Bioinformatics (Oxford, England)* 2005, **21**(21):4039-4045.

38. Katz MH: **Multivariable analysis: a primer for readers of medical research**. *Annals of internal medicine* 2003, **138**(8):644-650.

39. Wagner M, Naik D, Pothen A: **Protocols for disease classification from mass spectrometry data**. *Proteomics* 2003, **3**(9):1692-1698.

40. Fushiki T, Fujisawa H, Eguchi S: **Identification of biomarkers from mass spectrometry data using a "common" peak approach**. *BMC bioinformatics* 2006, **7**:358.

41. Wu B, Abbott T, Fishman D, McMurray W, Mor G, Stone K, Ward D, Williams K, Zhao H: **Comparison of statistical methods for classification of ovarian cancer using mass spectrometry data**. *Bioinformatics (Oxford, England)* 2003, **19**(13):1636-1643.

42. Diamandis EP: **Analysis of serum proteomic patterns for early cancer diagnosis: drawing attention to potential problems**. *Journal of the National Cancer Institute* 2004, **96**(5):353-356.

43. Lopez MF, Mikulskis A, Kuzdzal S, Bennett DA, Kelly J, Golenko E, DiCesare J, Denoyer E, Patton WF, Ediger R *et al*: **High-resolution serum proteomic profiling of Alzheimer disease samples reveals disease-specific, carrier-protein-bound mass signatures**. *Clinical chemistry* 2005, **51**(10):1946-1954.

44. Fisher WG, Rosenblatt KP, Fishman DA, Whiteley GR, Mikulskis A, Kuzdzal SA, Lopez MF, Tan NC, German DC, Garner HR: **A robust biomarker discovery pipeline for high-performance mass spectrometry data**. *Journal of bioinformatics and computational biology* 2007, **5**(5):1023-1045.

45. Hosmer DW, Lemeshow S: **Applied Logistic Regression**. New York: John Wiley & Sons, Inc.; 1989.

46. Yu CH: **An Overview of Remedial Tools for Collinearity in SAS**. In: *Proceedings of 2000 Western Users of SAS Software Conference: 2000*; 2000: 196-201.

47. Shtatland ES, Cain E, Barton MB: **The perils of stepwise logistic regression and how to escape them using information criteria and the Output Delivery System**. In: *SUGI'26 Proceedings: 2001*: Cary, NC: SAS Institute Inc.; 2001: Paper 222-226.

48. Breiman L, Friedman J, Olshen R, Stone C: **Classification and Regression Trees**. Belmont, CA: Wadsworth; 1984.

49. Eisen MB, Spellman PT, Brown PO, Botstein D: **Cluster analysis and display of genome-wide expression patterns**. *Proceedings of the National Academy of Sciences of the United States of America* 1998, **95**(25):14863-14868.

50. Datta S, Datta S: **Comparisons and validation of statistical clustering techniques for microarray gene expression data**. *Bioinformatics (Oxford, England)* 2003, **19**(4):459-466.

51. Rifai N, Gillette MA, Carr SA: **Protein biomarker discovery and validation: the long and uncertain path to clinical utility**. *Nature biotechnology* 2006, **24**(8):971-983.

52. Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M, Yasui Y: **Phases of biomarker development for early detection of cancer**. *Journal of the National Cancer Institute* 2001, **93**(14):1054-1061.

53. Zou KH, O'Malley AJ, Mauri L: **Receiver-operating characteristic analysis for evaluating diagnostic tests and predictive models**. *Circulation* 2007, **115**(5):654-657.

54. Dauvilliers Y, Arnulf I, Mignot E: **Narcolepsy with cataplexy**. *Lancet* 2007, **369**(9560):499-511.

55. Norden AG, Rodriguez-Cutillas P, Unwin RJ: **Clinical urinary peptidomics: learning to walk before we can run**. *Clin Chem* 2007, **53**(3):375-376.

56. Sheng S, Chen D, Van Eyk JE: **Multidimensional liquid chromatography separation of intact proteins by chromatographic focusing and reversed phase of the human serum proteome: optimization and protein database**. *Mol Cell Proteomics* 2006, **5**(1):26-34.

57. Lopez MF, Mikulskis A, Kuzdzal S, Golenko E, Petricoin EF, 3rd, Liotta LA, Patton WF, Whiteley GR, Rosenblatt K, Gurnani P *et al*: **A novel, high-throughput workflow for discovery and identification of serum carrier protein-bound peptide biomarker candidates in ovarian cancer samples**. *Clin Chem* 2007, **53**(6):1067-1074.

58. Haab BB: **Antibody arrays in cancer research**. *Mol Cell Proteomics* 2005, **4**(4):377-383.

59. Kostiainen R, Kotiaho T, Kuuranne T, Auriola S: **Liquid chromatography/atmospheric pressure ionization-mass spectrometry in drug metabolism studies**. *J Mass Spectrom* 2003, **38**(4):357-372.

60. Thornalley PJ, Battah S, Ahmed N, Karachalias N, Agalou S, Babaei-Jadidi R, Dawnay A: **Quantitative screening of advanced glycation endproducts in cellular and**

extracellular proteins by tandem mass spectrometry. *The Biochemical journal* 2003, **375**(Pt 3):581-592.

61. Roschinger W, Olgemoller B, Fingerhut R, Liebl B, Roscher AA: **Advances in analytical mass spectrometry to improve screening for inherited metabolic diseases**. *European journal of pediatrics* 2003, **162 Suppl 1**:S67-76.

62. Anderson L, Hunter CL: **Quantitative mass spectrometric multiple reaction monitoring assays for major plasma proteins**. *Mol Cell Proteomics* 2006, **5**(4):573-588.

63. Keshishian H, Addona T, Burgess M, Kuhn E, Carr SA: **Quantitative, multiplexed assays for low abundance proteins in plasma by targeted mass spectrometry and stable isotope dilution**. *Mol Cell Proteomics* 2007, **6**(12):2212-2229.

64. Anderson NL, Anderson NG, Haines LR, Hardie DB, Olafson RW, Pearson TW: **Mass spectrometric quantitation of peptides and proteins using Stable Isotope Standards and Capture by Anti-Peptide Antibodies (SISCAPA)**. *Journal of proteome research* 2004, **3**(2):235-244.

65. Berna MJ, Zhen Y, Watson DE, Hale JE, Ackermann BL: **Strategic use of immunoprecipitation and LC/MS/MS for trace-level protein quantification: myosin light chain 1, a biomarker of cardiac necrosis**. *Analytical chemistry* 2007, **79**(11):4199-4205.

66. Kulasingam V, Smith CR, Batruch I, Buckler A, Jeffery DA, Diamandis EP: **"Product ion monitoring" assay for prostate-specific antigen in serum using a linear ion-trap**. *Journal of proteome research* 2008, **7**(2):640-647.

67. Kuhn E, Wu J, Karl J, Liao H, Zolg W, Guild B: **Quantification of C-reactive protein in the serum of patients with rheumatoid arthritis using multiple reaction monitoring mass spectrometry and 13C-labeled peptide standards**. *Proteomics* 2004, **4**(4):1175-1186.

68. Barr JR, Maggio VL, Patterson DG, Jr., Cooper GR, Henderson LO, Turner WE, Smith SJ, Hannon WH, Needham LL, Sampson EJ: **Isotope dilution--mass spectrometric quantification of specific proteins: model application with apolipoprotein A-I**. *Clinical chemistry* 1996, **42**(10):1676-1682.

69. Wu SL, Amato H, Biringer R, Choudhary G, Shieh P, Hancock WS: **Targeted proteomics of low-level proteins in human plasma by LC/MSn: using human growth hormone as a model system**. *Journal of proteome research* 2002, **1**(5):459-465.

70. Barnidge DR, Goodmanson MK, Klee GG, Muddiman DC: **Absolute quantification of the model biomarker prostate-specific antigen in serum by LC-Ms/MS using protein cleavage and isotope dilution mass spectrometry**. *Journal of proteome research* 2004, **3**(3):644-652.

71. Cox DM, Zhong F, Du M, Duchoslav E, Sakuma T, McDermott JC: **Multiple reaction monitoring as a method for identifying protein posttranslational modifications**. *J Biomol Tech* 2005, **16**(2):83-90.

CHAPTER THREE
Case Study I: Multiple Sclerosis

3.1 INTRODUCTION

In this chapter, the acronym MS will refer exclusively to the disease Multiple Sclerosis and not the technique Mass Spectrometry.

3.1.1 Background

Multiple sclerosis (MS) is the most common inflammatory and demyelinating neurological disease of the brain. The prevalence varies from 1:600 to 1:2,000 with geographic location with incidences mostly in North America and Europe, affecting 2.5 million people worldwide and 400,000 in the United States alone [1] (Fig. 3.1). As with most autoimmune diseases, women are at a higher risk than men at a 3 to 1 ratio. MS manifests itself in young adulthood, predominantly between the ages of 20 and 40.

It is also debilitating with symptoms that include various degrees of paralysis, sensory disturbances, reduced coordination, and visual impairment such as optic neuritis. The molecular mechanism underlying the pathogenesis of MS remains unknown. MS is characterized by discrete areas of myelin, oligodendrocyte and axonal loss due to cellular and humoral immune responses responsible for severe inflammation and demyelination of the neurons. MS has long been regarded as a demyelinating disease but recent evidence suggests widespread axonal damage that correlates closely with the progression of disability [3]. It appears that MS is not just restricted in the white matter, but in the gray matter as well with occasional lack of lesions in white matter altogether.

The complexity of this disease is evident in the diagnosis stages, which consist of relapsing-remitting (RRMS), benign, primary progressive (PPMS), primary-remitting (PRMS) and secondary progressive (SPMS) (Fig. 3.2). Benign MS does not result in disability and the

patient returns to normal between attacks. Both RRMS and SPMS do not introduce new disability between attacks but the latter is followed by a steady increase in disability. PPMS patients see a steady increase in disability without attacks.

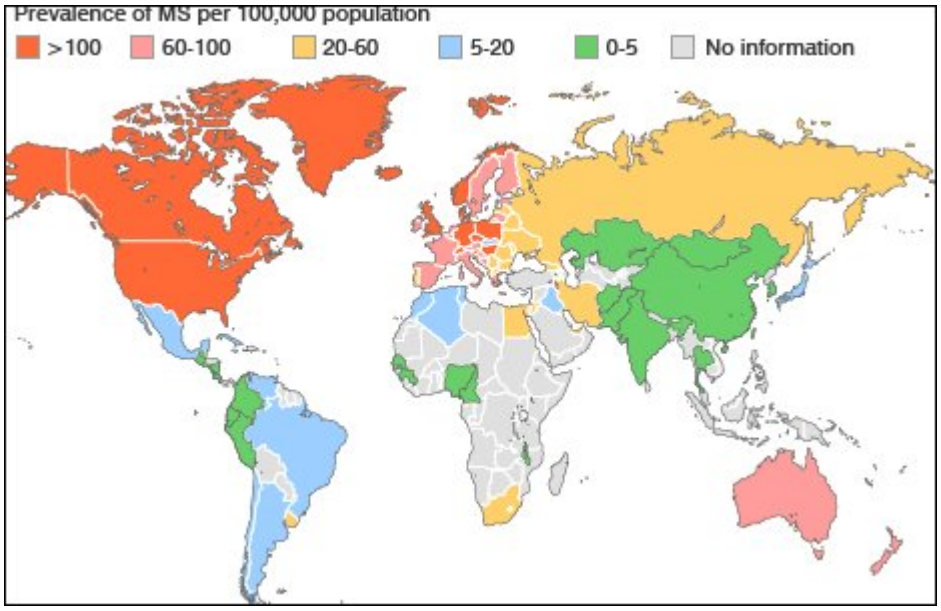


Figure 3.1 Worldwide prevalence of multiple sclerosis. (Internet source)

There are two distinct phases in MS: early (acute) and late (chronic). Relapses and disease exacerbations are vexing features of MS. Approximately 85% of MS is diagnosed in the acute RRMS stage [2], where inflammation dominates and new lesion sites appear in the white matter. During this stage, the patient undergoes episodic attacks and recovers, but with every attack, minor disability is incurred. Inflammation plays a major role in RRMS and it is at this stage where current therapies are most effective in controlling MS, mostly with anti-inflammatory drugs. Gradual progression to the chronic, fulminant SPMS causes global brain atrophy where lesions expand, leading to significant neurological disability and where inflammation purportedly decreases.

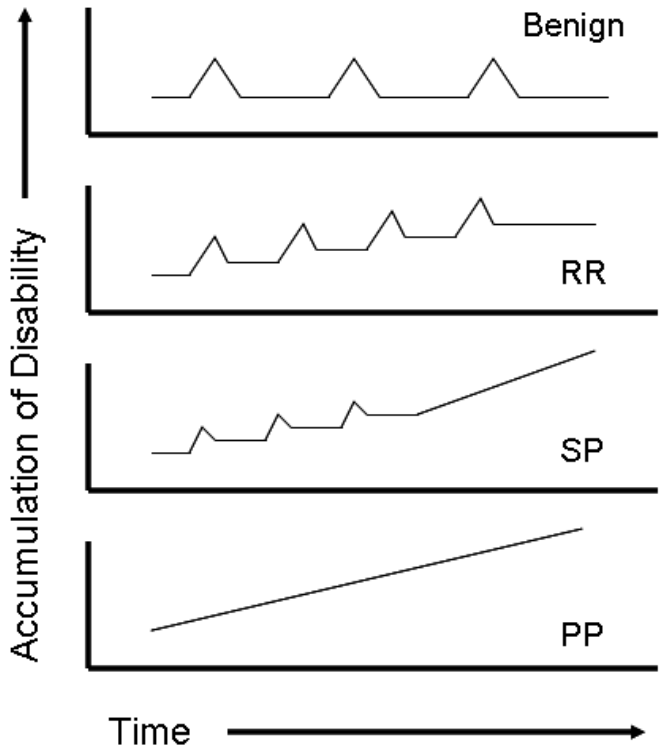


Figure 3.2 **Multifaceted forms of multiple sclerosis**. Spikes indicate episodic attacks.

The current model of MS pathogenesis involves genetic, environmental, and immunological factors. The susceptibility genes associated with MS are responsible for tipping the immunological balance by allowing T cells that recognize self-antigens to remain in circulation, escaping negative selection during maturation in the thymus. For the most part, this will not promote precipitation of the disease as they are kept at bay from their myelin-sheath derived antigens by an intact blood-brain barrier (BBB). However, an environmental factor such as a viral infection might compromise the integrity of this barrier and result in an inundation of autoreactive T cells into the brain, thereby initiating neuronal attacks and damages (Fig. 3.3).

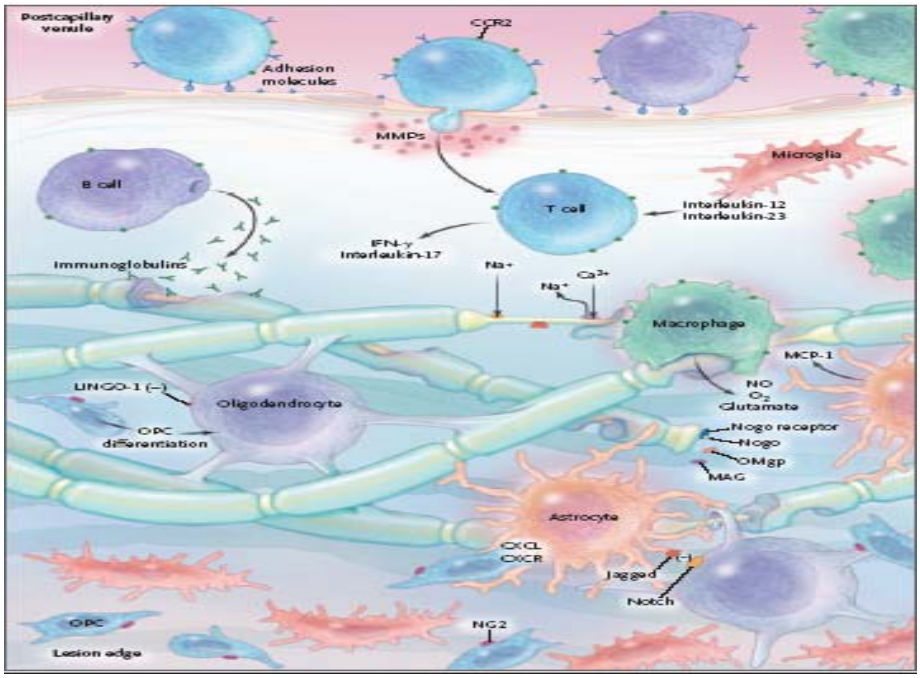


Figure 3.3 **Pathogenesis model of multiple sclerosis**. Myelin-reactive T cells in circulation are activated upon presentation to their antigen via molecular mimicry. Once they penetrate the BBB, they are reactivated by their target antigen and release toxic cytokines that result in neuronal damage. [3]

3.1.1.1 HLA associations

The heterogeneous pathogenesis of the lesions in MS is attributed to polygenic interactions notably in the HLA region [4, 5] (Table 3.1). These HLA genes are responsible for orchestrating the autoreactive immune cells in MS. Autoreactive $CD4^+$ T and B cells are inherently present but once activated, they target components of the myelin sheaths. The most common antigen is derived from the myelin basic protein, a 170 amino acid protein with 19 arginines and 12 lysines that contribute to it's basicity in interactions with negatively charged phosphate groups of membrane phospholipids to form a compact sheath. Subsequent release of myelin antigens promote a cascade of events that culminates in chronic disease development.

MHC class	Genetic associations	Odds ratio (increased risk of disease)
I	HLA-A*0301	1.9–2.1 (northern European Caucasoid populations)
II	HLA-DR2 haplotype (DRB1*1501–DRB5*0101–DQB1*0201)	2.9–3.6 (northern European Caucasoid populations)
	HLA-DR3 haplotype (DRB1*0301–DQA1*0501–DQB1*0201)	1.7 (Mediterranean populations)
	HLA-DR4 haplotype (DRB1*0405–DQA1*0501–DQB1*0301)	2.2 (Mediterranean populations)
	Alleles of the HLA-DPB1 locus (e.g. DPB1*0301)	1.6 (Mediterranean populations)

Table 3.1 Major histocompatibility complex (MHC) associations in multiple sclerosis.

3.1.2 Current diagnostic tools for multiple sclerosis

To date, there are no molecular tests for early detection of MS and diagnoses are currently based almost entirely by observation of clinical symptoms accompanied by CSF analysis, magnetic resonance imaging (MRI), and visual evoked potential (VEP) tests. Diagnosis of MS comes with elevated IgG index [6] and the presence of oligoclonal bands in the CSF, abnormalities (lesions) in the white matter via MRI, or delayed VEP [2]. These tests suffer from lack of disease sensitivity and specificity. Oligoclonal bands are not always present in MS patients and depending on the type of MRI used, some regions (subcortical or spinal cord) are not amenable to visualization. Furthermore, the readout parameters are not specific to MS. These methods are relied upon in the clinic simply because of the absence of a better test. Early clinical detection is considered late at the molecular level in MS since presentation of clinical symptoms indicates the disease has been present for a long time, and old lesions and some injury to the brain are already evident [3]. A more sensitive method to detect MS earlier or monitor its progress more accurately will dramatically improve the prospects of people who are predisposed to MS and serve as an impetus to the eventual replacement of clinical with molecular diagnosis.

Current therapies aim at reducing inflammation and slowing progression into the irreversible disability of SPMS. Thus, they are most effective at the RRMS stage and they include the likes of Natalizumab, Interferon Beta, Glatiramer acetate, and Methotrexate that inhibit T cell activation, proliferation, and migration across the BBB.

3.1.3 Proteomics studies on multiple sclerosis

Proteomic studies of MS have mostly involved 2DGE, either of human serum or CSF [7-9], with few studies using a high-throughput mass spectrometry-based platform for biomarker discovery [10, 11]. Avasarala *et al*. analyzed serum samples from 25 RRMS patients versus 25 healthy controls using MALDI TOF and found 3 *m/z* ratio peaks observed only in the MS samples [10]. Irani *et al*. analyzed CSF from 29 MS patients, 27 patients with transverse myelitis and 27 patients with other neurological diseases using SELDI TOF MS and suggested a cleavage product of Cystatin C to be a biomarker for a subgroup of MS patients [11]. A notable drawback of the first study was the use of healthy controls instead of patients with other inflammatory diseases, therefore bringing into question whether those 3 peaks are specific for MS. In the second study, the cleavage product of Cystatin C was subsequently shown to be an artifact of storage condition and not a specific marker for MS [12].

The limited number of reports on proteomic efforts in MS correlates with the infancy of proteomics and the technologies associated with this field. This suggests that the potential of biomarker discovery in MS still remains vastly untapped and more studies are warranted.

The following sections cover the first application of the high-throughput methodology described in Chapter 2 on the proteomic screening of the CSF proteome in a complex disease such as MS. The study is aimed at the discovery of potential biomarkers for MS to guide in the diagnosis and/or prognosis of the disease.

3.2 MATERIALS AND METHODS

3.2.1 Study population and source of CSF

CSF samples were analyzed from 34 patients diagnosed with MS and 26 patients with non-MS diagnoses (Appendix A). CSF samples were collected and centrifuged at 250g for 10 min to remove cellular debris. CSF supernatants were aliquoted and stored at -80°C until analysis. MS patients included 14 relapsing-remitting patients with a mean age of 38.1±10.5 years and age range of 22-60 years, and 20 secondary-progressive patients with a mean age of 45.1±11.7 years and age range of 28-67 years. Non-MS patients had a mean age of 60.4±2.4 years and age range of 26-91 years. Diagnoses of non-MS patients included prostate cancer (n=2), urinary/bladder cancer (n=1), congestive heart failure (n=2), headache (n=3), seizure (n=5), Parkinson's Disease (n=5), CNS tumors (n=5) and other neurological diseases (OND) (n=3). The RRMS and OND CSF samples were collected at UT Southwestern Medical Center under an IRB-approved protocol. The SPMS and other non-MS samples were acquired from the Human Brain and Spinal Fluid Resource Center at UCLA.

3.2.2 Sample preparation

CSF samples were aliquoted into 96-well microtiter plates and diluted 1:2 in 1X PBS, pH 7.4. All ProteinChip arrays were processed on the same day in a 96-well format as follows: IMAC30 ProteinChip arrays (Bio-Rad) from the same batch were activated with 50 mM nickel (II) sulfate hexahydrate three times for 15 mins with gentle shaking, followed by a quick rinse with HPLC grade water. The arrays were then equilibrated with 100 mL 1X PBS, pH 7.4 on a shaker at room temperature for 15 mins, thrice. The buffer was discarded and remaining droplets were aspirated from the spots using a vacuum tip. The IMAC30 arrays were then loaded into a ProteinChip

bioprocessor cassette to facilitate high-throughput analysis. The diluted CSF samples were deposited onto the array spots using a PerkinElmer MultiPROBE II PLUS HT EX liquid handler. Each CSF sample was run in triplicate. The bioprocessor cassette was sealed with aluminum foil, placed in a humidifying chamber, and incubated overnight at 4°C in a vacuum desiccator. The following day, the CSF samples were removed and the arrays were washed with 200 µL 1X PBS three times for 15 mins with gentle shaking at room temperature. The bioprocessor was subsequently removed and remaining droplets were aspirated from the spots using a vacuum tip. The ProteinChips were allowed to air dry for 15 mins. Then, two 1 µL aliquots of 5 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) matrix (LaserBio Labs, France) were added to each spot. All washing steps and matrix deposition were performed using a liquid handler to minimize operator bias. Matrix solution was kept protected from light at room temperature until ready for use. The spots were allowed to air dry before prOTOF MALDI-TOF mass spectrometry analysis.

3.2.3 MALDI TOF mass spectrometry analysis

ProteinChip arrays were placed in a custom made adapter for mass spectrometry analysis in the prOTOF2000 MALDI O-TOF mass spectrometer interfaced with TOFWorks software (PerkinElmer/SCIEX). Its orthogonal design enabled a single external mass calibrant to achieve better than 5 ppm mass accuracy over the 1,000 to 10,000 mass range acquired. A 2-point external calibration of the prOTOF instrument was performed before acquiring the spectra in a batch mode, four runs of six arrays at a time. Six spots were dedicated for the NIST reference serum sample and calibration was conducted at each run to ensure the integrity of the whole process. Acquisition was performed in one setting with a protocol optimized for ProteinChips: laser intensity of 78% at 100 Hz, 50V declustering voltage, 150 mL/min cooling flow rate, and 200 mL/min focusing flow rate. The prOTOF data files generated an average of 1 million data points per spectrum.

3.2.4 Biostatistical analysis

Raw spectra from the prOTOF were exported as text files using the prOTOF loader program and preprocessed to restore the zero-intensity values [13, 14]. The 60 CSF samples run in triplicate generated a total of 180 high-resolution mass spectra. Spectra from two CSF samples with macroscopic blood contamination were excluded from the analysis. The total ion current (TIC) of each spectrum was calculated and the average TIC was computed across the remaining 174 spectra. Spectra with a TIC value that was greater than twice or less than half of the average TIC were deemed outliers and were omitted from the study. This eliminated 4.6 to 6% of the spectra, depending on the group comparisons. Global normalization of the signal intensity of the mass peaks was performed by normalizing to the average TIC of the remaining spectra.

The mass spectral data set was analyzed both by an in-house *t*-test-based method [14] and Progenesis PG600 software (NonLinear Dynamics, UK) using Unweighted Pair Group Method with Arithmetic mean (UPGMA) hierarchical clustering. This serves to uncover consensus, differential mass peaks to reduce false positive associations to peaks that are due to overfitting or biases to a particular algorithm, as previously described in Section 2.4.2.

Analyses were performed to discover statistically differential markers between the following groups: MS vs non-MS, RRMS vs SPMS, MS vs PD, RRMS vs PD, SPMS vs PD, MS vs seizure and MS vs headache. The parameters for differential peak selection were set to include peaks with a minimum fold change of 20% between the groups compared and a p-value less than 0.05. The stringency of these parameters was tightened in certain comparisons to obtain peaks that are highly differential. ROC curve analysis was then performed using SAS (SAS Institute Inc., Cary, NC) on the differential peaks to determine their discriminatory power.

3.2.5 Western blot analysis

Samples from each group were separated by 10% Tris-Glycine-SDS-polyacrylamide gel electrophoresis and transferred to an Immobilon-PSQ 0.2 μm PVDF membrane (Millipore). Western blotting was performed on 9 μg total protein per lane. The membrane was blocked with 5% milk in 1X TBST overnight and then probed for 1 h with a mouse monoclonal antibody to Complement C3 (sc-52632, Santa Cruz Biotechnology Inc.) diluted 1:2,000. The membrane was washed with 1X TBST twice for 15 mins and then with 1X PBS twice for 15 mins, followed by incubation for 1 h with HRP-labeled goat anti-mouse secondary antibody (Bio-Rad) diluted 1:2,000. The washing steps were repeated before detection with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). The membrane was stripped, blocked and re-probed with rabbit polyclonal antibody against transthyretin (sc-13098, Santa Cruz Biotechnology Inc.) diluted 1:2,000 and HRP-labeled goat anti-rabbit secondary antibody (Bio-Rad) diluted 1:2,000. All antibodies were diluted in 5% milk in 1X TBST. All incubations were performed at room temperature. ImageJ was used for quantitation.

3.3 RESULTS

3.3.1 Analytical variables assessment

3.3.1.1 CSF dilution factor

The first parameter evaluated was the dilution ratio of samples to maximize binding of proteins/peptides to the capture surface. Previous experience with human serum samples run on the same platform showed that the optimal dilution factor was often 1:20. Since CSF has a much lower protein concentration than serum, the neat, 1:2 and 1:10 dilution factors were investigated. The neat and 1:2 samples produced similar mass spectra profiles with an equal number of peaks represented that were superior to the 10-fold diluted sample (Fig. 3.4). Therefore, a 1:2 dilution factor was adopted for the processing of all CSF samples to minimize sample consumption in the discovery phase and retain a majority of the same samples for downstream validation efforts.

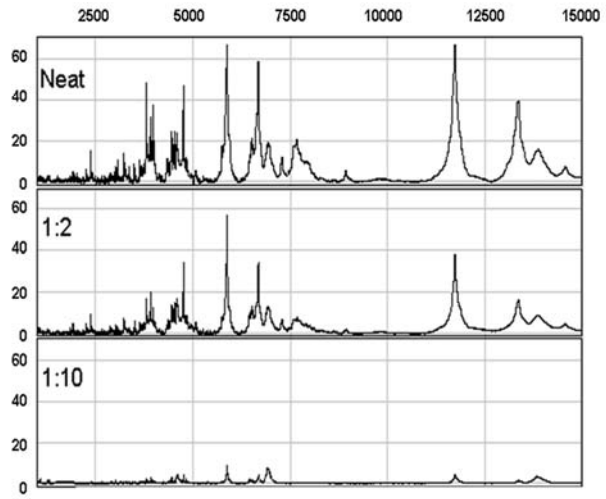


Figure 3.4 **Optimization of CSF dilution factor**. Mass spectra correspond to the same CSF sample analyzed either as neat, 2-fold diluted or 10-fold diluted on an IMAC30 ProteinChip. Signal intensity (y-axis) is plotted against the mass range, *m/z* (x-axis).

3.3.1.2 Surface retentate chemistry

As mentioned in Section 2.3.2, four main capture surface chemistries were evaluated for optimal peak observation and IMAC30 chips charged with nickel ions proved to be the best peak-producing platform. Thus, IMAC30 was the retentate surface of choice in the workflow.

3.3.1.3 Spectral reproducibility

As a check of spectral reproducibility, two CSF samples (one from each group being profiled) were run in triplicate. The high accuracy and resolution conferred by the prOTOF mass spectrometer enabled reproducible replicate spectra from individual samples to be obtained (Fig. 3.5).

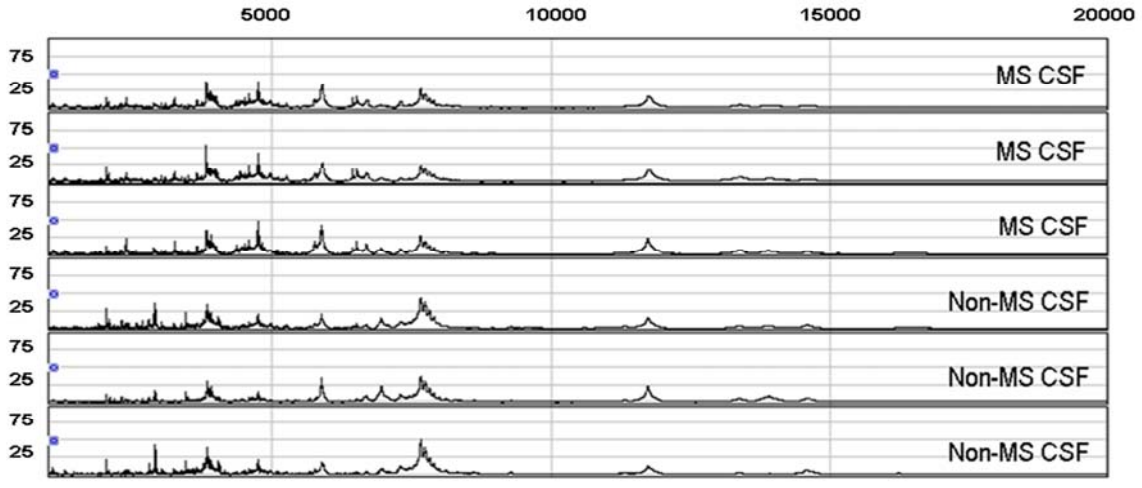
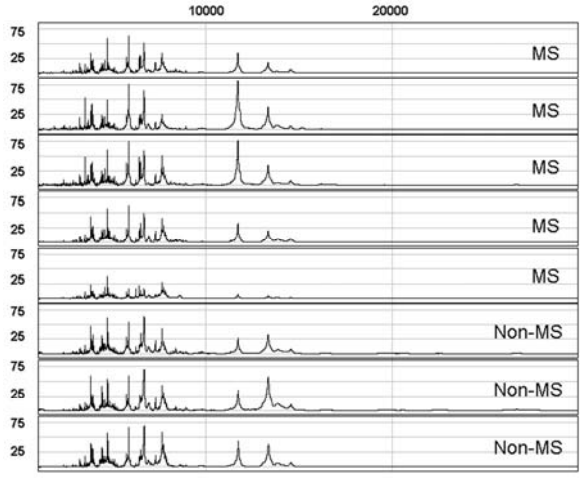


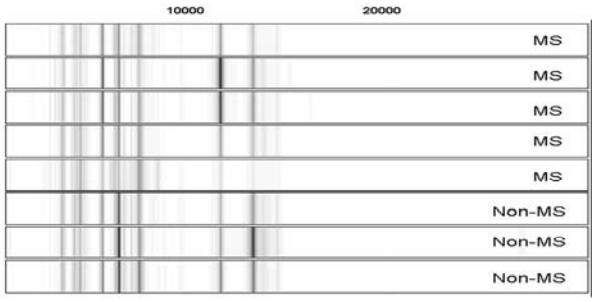
Figure 3.5 Spectral reproducibility was evaluated with an MS and a non-MS CSF sample run in triplicate on IMAC30 chip surface.

3.3.2 Pilot study

An initial pilot study was conducted to evaluate the methodology described in the previous chapter for detection of differential proteins in the CSF. This preliminary study involved CSF samples from MS (n=5) and non-MS (n=3) patients. Figure 3.6 shows that in addition to CSF being a protein-rich source of biological samples suitable for proteomic studies, there are detectable obvious differences in the protein profiles between MS and non-MS patients using this readout platform. Therefore, we proceeded to the larger MS profiling study of 60 samples.



A



B

Figure 3.6 **Preliminary study of multiple sclerosis**. Differential protein profiles between MS and non-MS samples were obtainable from CSF. The mass spectra (A) and corresponding gel views (B) are shown.

3.3.3 Profiling of multiple sclerosis and non-multiple sclerosis CSF

The workflow employed in this study is shown in Figure 3.7. The 60 clinical CSF samples were run in triplicate to minimize analytical variance in the technical process to generate a total of 180 mass spectra.

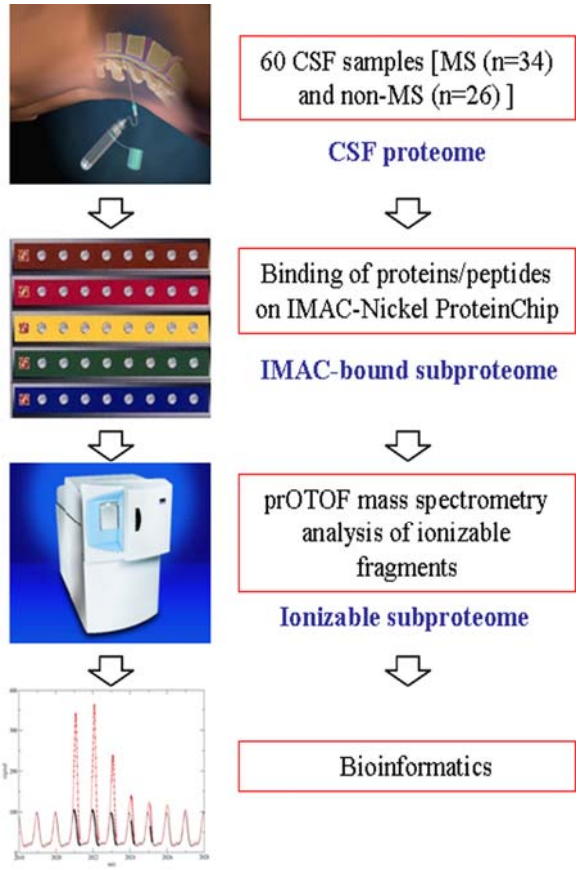


Figure 3.7 Cerebrospinal fluid-based biomarker discovery workflow

In this study, UPGMA hierarchical clustering and an in-house *t*-test-based method were used. Statistically significant differential peaks between the MS and non-MS groups are listed in Table 3.2.

In order to form the best discriminatory model for MS classification, consensus peaks that displayed the highest fold change between the MS and non-MS groups were selected for inclusion in the prediction model. To this end, mass peaks 2,091.91, 2,162.96, 2,294.00, and 2,898.47 were selected to form a biomarker panel collectively because they were discovered to be statistically differential with a fold change difference of at least two from the in-house t-test approach and were also selected as differential by the UPGMA algorithm. ROC curve analysis was performed on this four-peak model. The area under the curve (AUC) of 0.762 indicates good discriminatory power between the MS and non-MS groups (Fig. 3.8). In addition, this model has a sensitivity of 97%, specificity of 46%, and a positive predictive value (PPV) of 71% (Table 3.3).

Comparisons between MS to other non-MS diseases with at least three samples were also performed: MS vs PD, RR vs PD, SP vs PD, MS vs seizure, and MS vs headache. Comparisons to headache and seizure did not result in differential peaks. In the subgroup comparisons within MS (RRMS vs SPMS) and with Parkinson's Disease (PD) (MS vs PD, RRMS vs PD, SPMS vs PD), a collection of peaks were found to be differential with p-value less than 0.005 and a fold change of at least two (Table 3.4).

Mass peaks (*m/z*)	Fold change	P-value
UPGMA Hierarchical Clustering		
2,294.00	-1.56	<0.001
6,945.31	-1.53	0.038
2,898.47	-1.51	0.023
2,091.91	-1.49	<0.001
6,930.19	-1.46	0.033
2,162.96	-1.44	<0.001
6,962.27	-1.44	0.037
3,804.88	1.36	<0.001
2,861.34	-1.34	<0.001
2,880.42	-1.34	0.013
4,036.15	-1.28	<0.001
4,751.19	1.21	0.002
3,441.52	-1.21	0.009
T-test		
2,091.91	≥ 2.0	≤ 0.05
2,162.96	≥ 2.0	≤ 0.05
2,294.00	≥ 2.0	≤ 0.05
2,898.47	≥ 2.0	≤ 0.05
2,861.34	≥ 1.5	≤ 0.05
2,882.49	≥ 1.5	≤ 0.05
3,804.88	≥ 1.5	≤ 0.05

Table 3.2 **Differential peaks discovered from UPGMA hierarchical clustering and *t*-test for the comparison between the MS and non-MS groups**. Peaks are listed in decreasing fold change. Consensus peaks are in bold.

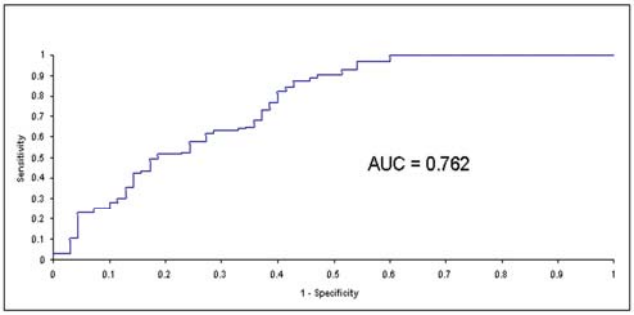


Figure 3.8 **ROC curve for MS versus non-MS model**. Model includes peaks at *m/z* 2,091, 2,162, 2,294, and 2,898.

ROC analysis of differential peaks between MS and non-MS

Sensitivity (%)	Specificity (%)	PPV (%)	Percent Accuracy (%)
96.91	45.71	71.21	75.45

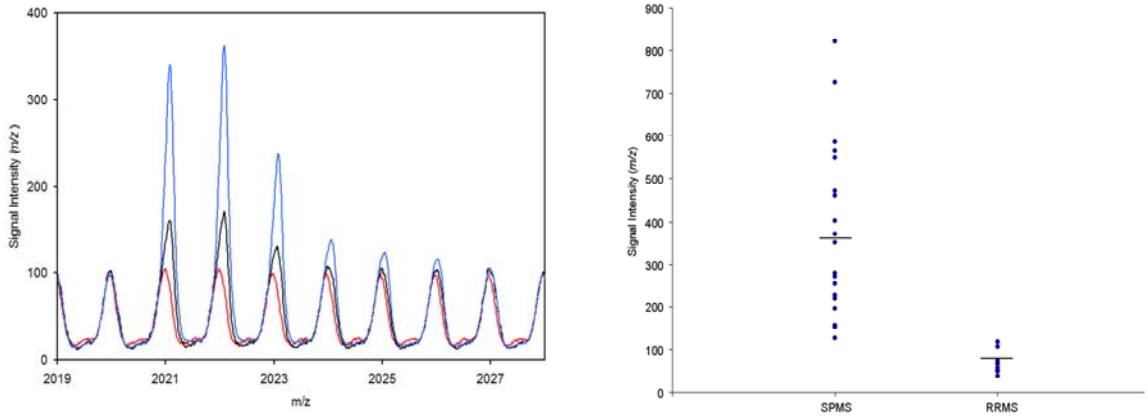
Table 3.3 **ROC curve analysis of differential peaks from group comparisons.** Diagnostic accuracy measures of MS versus non-MS model. Model includes peaks 2,091, 2,162, 2,294, and 2,898. PPV is the positive predictive value.

A peak at *m/z* = 2,021 was found to be 2.5-fold stronger in SPMS than RRMS. In fact, this differential peak was selectively present in the SPMS group but not in the RRMS group, as shown in Panels A and B of Figure 3.9 (p-value <0.0001). When the expression of this peak was interrogated in the PD CSF samples, it was also found to be elevated over the RRMS group, albeit at a lesser degree than the SPMS group (Panel A, Fig. 3.9). ROC curve analysis was performed on this peak to evaluate its discriminatory power between the three sample groups. It demonstrated strong discriminatory power between SPMS and RRMS (Fig. 3.10 and Table 3.5) with a high goodness of fit (Gof) score of 0.9117 and an AUC of 0.972. The sensitivity is 97% whereas the specificity is 88%. In contrast, this single-variate model with peak 2,021 has weak

discriminatory power between RRMS and PD (Gof score of 0.4039) and no discriminatory power between SPMS and PD (Gof score of 0.0477).

Group	RRMS vs SPMS	MS vs PD	SPMS vs PD	RRMS vs PD
m/z				
2,021.09	SPMS			
2,053.05	SPMS			
2,091.91		PD	PD	PD
2,162.96		PD	PD	PD
2,294.00		PD	PD	PD
3,804.88		MS	SPMS	RRMS

Table 3.4 **Differential peaks from subgroup comparisons**. The group where the peak was present at higher signal intensity is indicated. SPMS = secondary progressive MS, RRMS = relapsing-remitting MS, PD = Parkinson's disease.



A B

Figure 3.9 **Differential peak 2,021 between RRMS and SPMS**. (A) Overlay view of mean signal intensity of peak 2,021 in subgroup comparisons. Red trace, RRMS; blue trace, SPMS; black trace, PD (B) Peak 2,021 signal intensity in SPMS and RRMS CSF samples. Horizontal bars indicate the arithmetic mean.

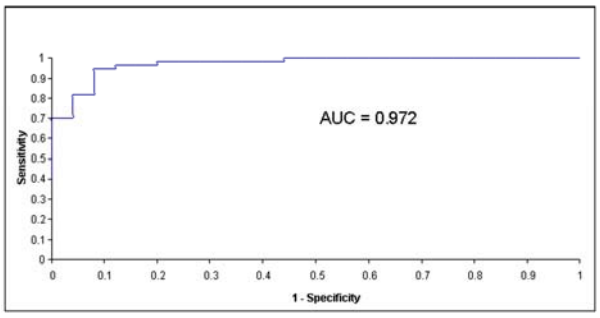


Figure 3.10 **ROC curve for RRMS versus SPMS model**. Model includes peak at *m/z* 2,021.

ROC analysis of the differential peak 2,021 between RRMS and SPMS

Sensitivity (%)	Specificity (%)	PPV (%)	Percent Accuracy (%)
96.67	88.00	95.08	94.11

Table 3.5 **ROC curve analysis of differential peaks from subgroup comparison.** Diagnostic accuracy measures of RRMS versus SPMS model. PPV is the positive predictive value.

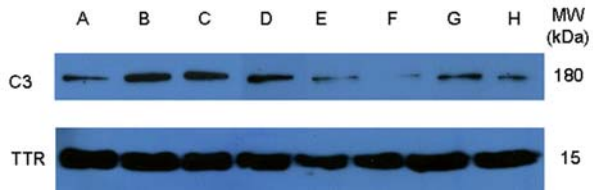
3.3.4 Identification of the 2,021 peak as a Complement C3 fragment

Significant efforts to scale up the enrichment protocol and obtain a sample suitable for analysis by tandem FT-MS were unsuccessful. Therefore, we turned to *in silico* methods to generate a hypothesis for the molecular identity of the 2,021 mass peak. A detailed literature search of proteins identified in human CSF to date revealed that a peak of *m/z* = 2,021 had been identified previously as a fragment of the Complement C3 protein [15], called the C3f proteolytic fragment. Coincidentally, in a separate pattern profiling study employing human serum, rather than CSF, samples from PD and non-PD patients, a peak with *m/z* = 2,021 bound to serum albumin was found to be elevated in the PD samples. This differential peak was identified to be the C3f

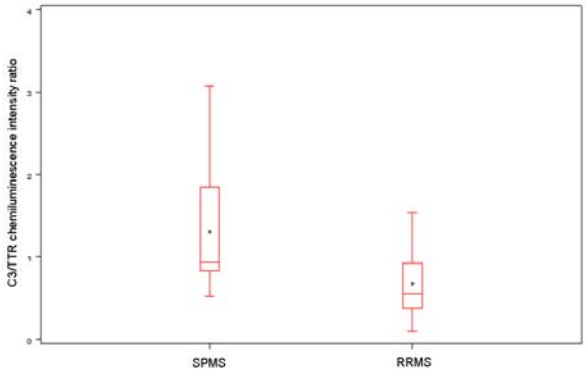
fragment of the Complement C3 precursor (NCBI database search by FT-MS) [16, 17]. These observations, and the fact that our peak 2,021 is also elevated in the PD CSF samples, prompted us to determine whether our 2,021 peak might correspond to the C3f fragment in CSF.

We initially attempted to probe this point at the peptide level by immunoprecipitating the peptide corresponding to the 2,021 peak from CSF using immobilized antibodies that recognize an epitope in C3f on ProteinChips, but this proved unsuccessful. Therefore, we hypothesized that if the 2,021 peak is the Complement C3f fragment, then perhaps the level of intact Complement C3 protein might also be diagnostic and so proceeded to address this possibility.

We analyzed the level of intact Complement C3 in both the RRMS and SPMS subgroups in CSF by Western blot analysis. The Complement C3 protein levels were ascertained with an antibody that recognizes an epitope in the C3f fragment. 9 μg of total CSF protein was loaded for each sample on a 10% gel. Transthyretin (TTR) was probed as a loading control for CSF. All remaining samples that were not consumed in prior identification efforts (16 from SPMS and 8 from RRMS) were run and probed for Complement C3 concurrently to minimize technical bias. It was found that the expression of Complement C3 was elevated in the SPMS CSF samples over RRMS (Panel A, Fig. 3.11). The chemiluminescence intensity was quantified and the ratio of C3 intensity over TTR intensity was calculated. The C3/TTR ratio between the RRMS and SPMS groups showed a statistically significant difference with a p-value of <0.05. The box and whisker plot of the ratios are shown in Panel B of Figure 3.11. The protein was more abundant in the CSF of SPMS patients compared with RRMS cases. Based on this tight correlation between the Western blot data for Complement C3 protein and the mass spectrometry data, we propose that the m/z = 2,021 peak identified as a possible biomarker is the C3f fragment of Complement C3, though we cannot completely rule out the possibility that the 2,021 peak represents a different molecule whose level is somehow linked tightly to the level of Complement C3.



A



B

Figure 3.11 **Differential levels of Complement C3 in the CSF of RRMS and SPMS patients**. (A) Representative Western blot depicting stage-specific expression of Complement C3. A-D= SPMS CSF, E-H=RRMS CSF. TTR was used as loading control. (B) Box and whisker plot of C3/TTR chemiluminescence intensity ratio.

3.4 DISCUSSION

CSF is produced in the choroid plexus of the brain and is replaced three to four times a day, reflecting the current physiological state of the central nervous system (CNS). Because of this, and its close proximity to the lesions that characterize MS, this biological fluid is likely to be a rich source of highly concentrated CNS-related disease protein biomarkers. Furthermore, the presence of a protein marker in proximal fluid may be a surrogate for its availability to the systemic circulation and can thus be probed in blood for clinical assays. In addition, CSF can be procured in a relatively non-invasive procedure and even though more than 80% of CSF proteins

originate from plasma [18], it is a less complex proteome analyze because of its lower dynamic range in protein concentration (10^8) than plasma (10^{10}) [19]. The CSF samples were not pre-fractionated due to its lower protein content.

In this study, we searched for interesting protein biomarkers in the CSF of MS patients by coupling surface retentate chromatography, using ProteinChips, to an accurate, high-performance mass spectrometry readout. To date, only limited studies involving CSF based on this high-throughput platform have been reported. Our previous study [16] and those done by others [20] have shown that the number of output peaks in the mass spectra is highly dependent on the sample processing and data acquisition conditions. Therefore, the optimization of several parameters for CSF analysis was undertaken prior to the actual profiling study.

In this report, we generated and analyzed 180 CSF proteome profiles from 60 patients to discover biomarkers specific for MS. In the MS versus non-MS comparison, four major peaks with the largest number of discriminating data points [14] were incorporated into a prediction model to assess their ability to classify samples from both groups. The area under the ROC curve of 0.762 shows that these four peaks have good discriminatory power between MS and non-MS samples. They exhibit a sensitivity of 97% and a specificity of 46%. Even though the specificity in this comparison is quite low, any molecular diagnostics with specificity significantly greater than zero is of interest in MS as there is currently no molecular test available.

The strength of this study lies in the heterogeneity of both the clinical MS and non-MS control CSF samples that encompassed patients from different stages of MS and a battery of other neurological diseases. It is advantageous if the sample used in a study is representative of samples that will be procured in the future for application in the same workflow. The inclusion of CSF samples from healthy people in the study is meaningless as healthy people in general will not have the incentive to undergo spinal taps or testing for MS. In addition, we reasoned that a control group consisting of CSF samples from patients with other non-MS diseases, especially those of inflammatory origin, would confer a specificity for MS onto the resulting candidate

biomarkers. Moreover, it lends itself to subgroup comparisons. CSF samples from the two major stages of MS allowed us to perform pattern profiling comparison of samples within the same neurodegenerative disease (RRMS vs SPMS) to uncover differential mass peaks to aid in the subclassification of MS. Also of interest to us was the five CSF samples procured from PD patients in the non-MS group. This enabled us to perform differential peak discovery across neurodegenerative diseases (MS vs PD, RRMS vs PD, SPMS vs PD).

The correlation between newly discovered biomarkers and the disease is usually lower during validation. This inconsistency can be reduced by increasing specificity in the discovery phase through the incorporation of disease pathway knowledge, such as the progression from RRMS to SPMS. Approximately 85% of early phase MS is diagnosed in the acute, RRMS stage where inflammation is known to dominate and drug intervention is still feasible. On the contrary, progression to the chronic, SPMS stage could lead to significant accumulation of disability and is a situation where drug options are limited. Therefore, any diagnostic marker that could discriminate between these two stages of MS would be useful in monitoring the progression of the disease. In particular, peak 2,021 drew our attention because (i) of all the differential peaks for this RRMS vs SPMS subgroup comparison, it was the only peak to display a greater than 2-fold change between the groups across both statistical platforms and (ii) in our in-house *t*-test approach, which provides a weighting factor to each differential peak, this peak had the most discriminating data points. In addition, the peak displayed strong signal intensity with a high signal-to-noise ratio, making it an excellent candidate for downstream identification via tandem mass spectrometry sequencing. Determining the identity of this peak will possibly shed some light on the process that underlies the neurodegenerative process MS.

In the field of proteomic biomarker discovery, enriching for low abundance differential peaks from complex biological samples for identification remains a challenging endeavor. This is especially true in our case where our differential signature peaks are less than 10,000 Da and are dominated by peptides or small protein fragments. This makes it difficult to resolve the species on

a gel from the complex CSF proteome as an enrichment strategy for peptide mass fingerprinting via in-gel trypsin digestion. Initial attempts to simplify the proteome and enrich for this peak for sequencing via orthogonal chromatographic separations did not result in positive identifications. Therefore, we next performed an extensive literature search of CSF proteins identified to date via mass spectrometry and found that a peak with *m/z* = 2,021 had previously been identified as the C3f fragment of Complement C3 [15]. Interestingly, in a parallel and unrelated pattern profiling study of human serum samples from PD and non-PD patients in one of our laboratories [16], a molecule with *m/z* = 2,021 bound to albumin was also found to be elevated in the PD samples and subsequently identified by FT-MS to be a fragment of the Complement C3 precursor [17]. It is a striking coincidence that in both profiling studies of neurodegenerative diseases, a peak at mass 2,021 was found to be differential and elevated in the group with more severe neuronal damage (SPMS and PD). Furthermore, the same mass peak was found to be elevated in the PD samples, irrespective of biological source.

It is conceivable that the C3f fragment bound to albumin found in the serum study is also present in the CSF because (i) approximately 80% of CSF proteins are derived from blood [18] with albumin dominating the CSF proteome as it does in plasma, (ii) since albumin is only synthesized in the liver, its presence in the CSF must originate from blood [19], and (iii) leakage of molecular species of brain origin to the blood and vice versa has been documented [21, 22]. This led us to hypothesize that the peak 2,021 found in our CSF study could be of the same molecule found in both the CSF and serum studies of neurological diseases aforementioned.

The CSF samples from both MS subgroups were probed with an antibody whose epitope includes the C3f fragment sequence. The relative level of the Complement C3 protein was found to correlate well to that of the 2,021 peak in our proteomic analysis, with higher expression level in the SPMS group. The difference between the RRMS and SPMS groups is statistically significant with a p-value of <0.05 over 16 SPMS and 8 RRMS individual CSF samples. TTR, a pre-albumin that is abundant in the CSF, was used as a loading control. Given these data, we

believe that the molecule identified in our study with *m/z* = 2,021 is the C3f fragment of Complement C3, but since we have not identified this species directly, we cannot completely dismiss the unlikely possibility that it is a different molecule with the same mass whose level is correlated with the level of Complement C3 protein. Efforts to unequivocally identify the differential 2,021 peak from CSF are ongoing in our laboratory. This subgroup comparison is admittedly limited in size as a *post hoc* analysis for the difference observed in the Western blots suggested that at least 23 samples per group being compared will have to be represented in order to achieve a desired power of 90%.

Nonetheless, the discovery that the level of intact Complement C3 protein appears to be a useful marker for MS should make it easier to carry out more extensive future studies of far more patient samples. Indeed, it should be possible to design an ELISA assay for the level of this protein in the CSF and perhaps for the C3f fragment as well, allowing the facile measurement of the ratio of these polypeptides, which may be useful.

Of course, the data presented here merely support a correlation between the level of Complement C3 and the C3f fragment with MS and do not prove a causal relationship. Nonetheless, it would not be surprising if this were indeed the case. The complement system, as part of the innate immune system, is involved in a repertoire of activities that include the recognition and killing of pathogens, and the initiation of an inflammatory response via the bioactive fragments that are generated during activation and their subsequent degradation. Complement has long been implicated in the pathogenesis of MS and its animal model EAE [23-26]. Cytotoxic complement components can be present in the brain parenchyma either through transudation through a compromised BBB [27], or from local expression by resident brain cells [28, 29]. Even though complement proteins are usually at a low level in the CSF, complement expression can be induced by antibodies against myelin proteins found in MS patients [30], by myelin and oligodendrocytes themselves [31, 32], or by cytokines and viruses [33, 34]. Neurons and oligodendrocytes are especially susceptible to complement-mediated injury because although

they are capable of producing complement proteins and complement anaphylatoxin receptors, unlike astrocytes and microglia, they are deficient in membrane complement regulators [35, 36], contributing to their selective damage in MS [37].

C3 and C5 are the major components of complement with the ability to produce anaphylatoxins as inflammation mediators. In stark contrast to C5, C3 has been shown to be required for maximal disease progression to the chronic stage in a protein dose-dependent manner in EAE [25, 26, 38], possibly through its regulatory role of encephalitogenic T cells [39]. Elevated C3 activity has been reported in the serum or CSF of MS patients experiencing an acute relapse or secondary progressive disease [40].

Complement C3 has been implicated as a potential biomarker of MS in general [41] with recent mass spectrometry-based studies identifying Complement C3 as a differential protein in the MS over non-MS CSF samples [42, 43]. However, our data suggest that it may be adopted as a stage-specific marker, given it was found to be present at a higher level in the CSF of SPMS patients when probed across 16 SPMS and 8 RRMS CSF samples. Moreover, the elevated expression of C3 in SPMS over RRMS correlated extremely well with the differential peak 2,021 whose putative identity from database search is the complement C3 fragment, C3f.

The initial activation of C3 produces the cleavage products C3a and C3b. C3a is an anaphylotoxin that serves as an inflammation mediator involved in chemotaxis, smooth muscle contractions and increased vascular permeability. C3b and its proteolytic fragments serve as opsonins that enhance phagocytosis and lymphocyte activation [25, 44]. Endogenous proteolytic enzymes can subsequently cleave C3b into C3bi and C3f. C3bi is a membrane-bound intermediate whereas C3f is a free, diffusible component. Little is known about the physiologic role of C3f to date but C3f appears to be functionally related to C3a since it is also a spasmogenic factor and is able to enhance vascular permeability at physiologic levels [45], suggesting its role in inflammation via enhanced vascular permeability to promote leukocyte diapedesis. Furthermore, C3f potentially interacts with the C3a receptor [45]. Thus, the rapid release of C3f

may result in lysis of neuronal cells in the vicinity, promoting autoimmune demyelination to occur.

3.5 CONCLUSION

We have incorporated a highly-reproducible, high-throughput proteomic platform for the analysis of clinical samples to obtain a panel of candidate proteins for diagnosis and staging of MS. CSF represents an ideal biological source for CNS disease studies as it is routinely drawn for diagnostic purposes and is rich in CNS-related biomarkers. We propose Complement C3 as a CSF biomarker indicative of disease progression to address the paucity of diagnostic molecular biomarkers in MS. Even though our validation set is admittedly limited in size and validation in a larger sample set is warranted, we envision C3 could be developed into a clinical test that would be amenable for routine prognostic evaluations. Furthermore, it is anticipated that complement therapeutics that target either the C3 protein or C3 convertase which generates the activated C3 fragments will significantly attenuate the disease, as seen in EAE. A clinical outcome of even a modest delay in progression of MS would lead to a greater quality of life than that resulting from any currently available treatment.

3.6 BIBLIOGRAPHY

1. Shoulson I: **Experimental therapeutics of neurodegenerative disorders: unmet needs**. *Science (New York, NY* 1998, 282(5391):1072-1074.
2. Murray TJ: **Diagnosis and treatment of multiple sclerosis**. *Bmj* 2006, 332(7540):525-527.
3. Frohman EM, Racke MK, Raine CS: **Multiple sclerosis--the plaque and its pathogenesis**. *The New England journal of medicine* 2006, 354(9):942-955.
4. Yeo TW, De Jager PL, Gregory SG, Barcellos LF, Walton A, Goris A, Fenoglio C, Ban M, Taylor CJ, Goodman RS *et al*: **A second major histocompatibility complex susceptibility locus for multiple sclerosis**. *Ann Neurol* 2007, 61(3):228-236.
5. Dyment DA, Herrera BM, Cader MZ, Willer CJ, Lincoln MR, Sadovnick AD, Risch N, Ebers GC: **Complex interactions among MHC haplotypes in multiple sclerosis: susceptibility and resistance**. *Hum Mol Genet* 2005, 14(14):2019-2026.
6. Lefranc D, Almeras L, Dubucquoi S, de Seze J, Vermersch P, Prin L: **Distortion of the self-reactive IgG antibody repertoire in multiple sclerosis as a new diagnostic tool**. *J Immunol* 2004, 172(1):669-678.
7. Almeras L, Lefranc D, Drobecq H, de Seze J, Dubucquoi S, Vermersch P, Prin L: **New antigenic candidates in multiple sclerosis: identification by serological proteome analysis**. *Proteomics* 2004, 4(7):2184-2194.
8. Dumont D, Noben JP, Raus J, Stinissen P, Robben J: **Proteomic analysis of cerebrospinal fluid from multiple sclerosis patients**. *Proteomics* 2004, 4(7):2117-2124.
9. Hammack BN, Fung KY, Hunsucker SW, Duncan MW, Burgoon MP, Owens GP, Gilden DH: **Proteomic analysis of multiple sclerosis cerebrospinal fluid**. *Multiple sclerosis (Houndmills, Basingstoke, England)* 2004, 10(3):245-260.

10. Avasarala JR, Wall MR, Wolfe GM: **A distinctive molecular signature of multiple sclerosis derived from MALDI-TOF/MS and serum proteomic pattern analysis: detection of three biomarkers**. *J Mol Neurosci* 2005, 25(1):119-125.

11. Irani DN, Anderson C, Gundry R, Cotter R, Moore S, Kerr DA, McArthur JC, Sacktor N, Pardo CA, Jones M *et al*: **Cleavage of cystatin C in the cerebrospinal fluid of patients with multiple sclerosis**. *Ann Neurol* 2006, 59(2):237-247.

12. Del Boccio P, Pieragostino D, Lugaresi A, Di Ioia M, Pavone B, Travaglini D, D'Aguanno S, Bernardini S, Sacchetta P, Federici G *et al*: **Cleavage of cystatin C is not associated with multiple sclerosis**. *Annals of neurology* 2007, 62(2):201-204; discussion 205.

13. Lopez MF, Mikulskis A, Kuzdzal S, Bennett DA, Kelly J, Golenko E, DiCesare J, Denoyer E, Patton WF, Ediger R *et al*: **High-resolution serum proteomic profiling of Alzheimer disease samples reveals disease-specific, carrier-protein-bound mass signatures**. *Clinical chemistry* 2005, 51(10):1946-1954.

14. Fisher WG, Rosenblatt KP, Fishman DA, Whiteley GR, Mikulskis A, Kuzdzal SA, Lopez MF, Tan NC, German DC, Garner HR: **A robust biomarker discovery pipeline for high-performance mass spectrometry data**. *Journal of bioinformatics and computational biology* 2007, 5(5):1023-1045.

15. Jimenez CR, Koel-Simmelink M, Pham TV, van der Voort L, Teunissen CE: **Endogeneous peptide profiling of cerebrospinal fluid by MALDI-TOF mass spectrometry: Optimization of magnetic bead-based peptide capture and analysis of preanalytical variables**. *Proteomics Clinical applications* 2007, 1:1385-1392.

16. German DC, Gurnani P, Nandi A, Garner HR, Fisher W, Diaz-Arrastia R, O'Suilleabhain P, Rosenblatt KP: **Serum biomarkers for Alzheimer's disease: proteomic discovery**. *Biomedicine & pharmacotherapy = Biomedecine & pharmacotherapie* 2007, 61(7):383-389.

17. Lopez MF, Mikulskis A, Kuzdzal S, Golenko E, Petricoin EF, 3rd, Liotta LA, Patton WF, Whiteley GR, Rosenblatt K, Gurnani P *et al*: **A novel, high-throughput workflow for discovery and identification of serum carrier protein-bound peptide biomarker candidates in ovarian cancer samples**. *Clinical chemistry* 2007, 53(6):1067-1074.

18. Thompson EJ, Keir G: **Laboratory investigation of cerebrospinal fluid proteins**. *Annals of clinical biochemistry* 1990, 27 (Pt 5):425-435.

19. Blennow K, Fredman P, Wallin A, Gottfries CG, Karlsson I, Langstrom G, Skoog I, Svennerholm L, Wikkelso C: **Protein analysis in cerebrospinal fluid. II. Reference values derived from healthy individuals 18-88 years of age**. *European neurology* 1993, 33(2):129-133.

20. Guerreiro N, Gomez-Mancilla B, Charmont S: **Optimization and evaluation of surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry for protein profiling of cerebrospinal fluid**. *Proteome science* 2006, 4:7.

21. Shibata M, Yamada S, Kumar SR, Calero M, Bading J, Frangione B, Holtzman DM, Miller CA, Strickland DK, Ghiso J *et al*: **Clearance of Alzheimer's amyloid-ss(1-40) peptide from brain by LDL receptor-related protein-1 at the blood-brain barrier**. *The Journal of clinical investigation* 2000, 106(12):1489-1499.

22. Tang W, Duke-Cohan JS: **Human secreted attractin disrupts neurite formation in differentiating cortical neural cells in vitro**. *Journal of neuropathology and experimental neurology* 2002, 61(9):767-777.

23. Sellebjerg F, Jaliashvili I, Christiansen M, Garred P: **Intrathecal activation of the complement system and disability in multiple sclerosis**. *Journal of the neurological sciences* 1998, 157(2):168-174.

24. Storch MK, Piddlesden S, Haltia M, Iivanainen M, Morgan P, Lassmann H: **Multiple sclerosis: in situ evidence for antibody- and complement-mediated demyelination**. *Annals of neurology* 1998, 43(4):465-471.

25. Barnum SR, Szalai AJ: **Complement and demyelinating disease: no MAC needed?** *Brain research reviews* 2006, 52(1):58-68.

26. Szalai AJ, Hu X, Adams JE, Barnum SR: **Complement in experimental autoimmune encephalomyelitis revisited: C3 is required for development of maximal disease**. *Molecular immunology* 2007, 44(12):3132-3136.

27. Lindsberg PJ, Ohman J, Lehto T, Karjalainen-Lindsberg ML, Paetau A, Wuorimaa T, Carpen O, Kaste M, Meri S: **Complement activation in the central nervous system following blood-brain barrier damage in man**. *Annals of neurology* 1996, 40(4):587-596.

28. Morgan BP, Gasque P: **Expression of complement in the brain: role in health and disease**. *Immunology today* 1996, 17(10):461-466.

29. Thomas A, Gasque P, Vaudry D, Gonzalez B, Fontaine M: **Expression of a complete and functional complement system by human neuronal cells in vitro**. *International immunology* 2000, 12(7):1015-1023.

30. Reindl M, Linington C, Brehm U, Egg R, Dilitz E, Deisenhammer F, Poewe W, Berger T: **Antibodies against the myelin oligodendrocyte glycoprotein and the myelin basic protein in multiple sclerosis and other neurological diseases: a comparative study**. *Brain* 1999, 122 (Pt 11):2047-2056.

31. Vanguri P, Koski CL, Silverman B, Shin ML: **Complement activation by isolated myelin: activation of the classical pathway in the absence of myelin-specific antibodies**. *Proceedings of the National Academy of Sciences of the United States of America* 1982, 79(10):3290-3294.

32. Wren DR, Noble M: **Oligodendrocytes and oligodendrocyte/type-2 astrocyte progenitor cells of adult rats are specifically susceptible to the lytic effects of complement in absence of antibody.** *Proceedings of the National Academy of Sciences of the United States of America* 1989, 86(22):9025-9029.

33. Argaw AT, Zhang Y, Snyder BJ, Zhao ML, Kopp N, Lee SC, Raine CS, Brosnan CF, John GR: **IL-1beta regulates blood-brain barrier permeability via reactivation of the hypoxia-angiogenesis program**. *J Immunol* 2006, 177(8):5574-5584.

34. Bruder C, Hagleitner M, Darlington G, Mohsenipour I, Wurzner R, Hollmuller I, Stoiber H, Lass-Florl C, Dierich MP, Speth C: **HIV-1 induces complement factor C3 synthesis in astrocytes and neurons by modulation of promoter activity**. *Molecular immunology* 2004, 40(13):949-961.

35. Singhrao SK, Neal JW, Rushmere NK, Morgan BP, Gasque P: **Spontaneous classical pathway activation and deficiency of membrane regulators render human neurons susceptible to complement lysis**. *The American journal of pathology* 2000, 157(3):905-918.

36. Koski CL, Estep AE, Sawant-Mane S, Shin ML, Highbarger L, Hansch GM: **Complement regulatory molecules on human myelin and glial cells: differential expression affects the deposition of activated complement proteins**. *Journal of neurochemistry* 1996, 66(1):303-312.

37. Mead RJ, Neal JW, Griffiths MR, Linington C, Botto M, Lassmann H, Morgan BP: **Deficiency of the complement regulator CD59a enhances disease severity, demyelination and axonal injury in murine acute experimental allergic encephalomyelitis**. *Laboratory investigation; a journal of technical methods and pathology* 2004, 84(1):21-28.

38. Boos L, Campbell IL, Ames R, Wetsel RA, Barnum SR: **Deletion of the complement anaphylatoxin C3a receptor attenuates, whereas ectopic expression of C3a in the brain exacerbates, experimental autoimmune encephalomyelitis**. *J Immunol* 2004, 173(7):4708-4714.

39. Kemper C, Atkinson JP: **T-cell regulation: with complements from innate immunity**. *Nature reviews* 2007, 7(1):9-18.

40. Rus H, Cudrici C, David S, Niculescu F: **The complement system in central nervous system diseases**. *Autoimmunity* 2006, 39(5):395-402.

41. Barnum SR, Szalai AJ: **Complement as a biomarker in multiple sclerosis**. *Journal of neuropathology and experimental neurology* 2005, 64(8):741.

42. Noben JP, Dumont D, Kwasnikowska N, Verhaert P, Somers V, Hupperts R, Stinissen P, Robben J: **Lumbar cerebrospinal fluid proteome in multiple sclerosis: characterization by ultrafiltration, liquid chromatography, and mass spectrometry**. *Journal of proteome research* 2006, 5(7):1647-1657.

43. Stoop MP, Dekker LJ, Titulaer MK, Burgers PC, Sillevis Smitt PA, Luider TM, Hintzen RQ: **Multiple sclerosis-related proteins identified in cerebrospinal fluid by advanced mass spectrometry**. *Proteomics* 2008, 8(8):1576-1585.

44. Prineas JW, Kwon EE, Cho ES, Sharer LR, Barnett MH, Oleszak EL, Hoffman B, Morgan BP: **Immunopathology of secondary-progressive multiple sclerosis**. *Annals of neurology* 2001, 50(5):646-657.

45. Ganu VS, Muller-Eberhard HJ, Hugli TE: **Factor C3f is a spasmogenic fragment released from C3b by factors I and H: the heptadeca-peptide C3f was synthesized and characterized**. *Molecular immunology* 1989, 26(10):939-948.

CHAPTER FOUR
Case Study II: Narcolepsy

4.1 INTRODUCTION

4.1.1 Background

Narcolepsy is a neurological disorder known to affect sleep states. Albeit non-life threatening, it is a debilitating and lifelong neurologic disease which impacts the daily lives of individuals severely since a person with narcolepsy is likely to become drowsy or to fall asleep during the day, often at inappropriate times and places. It is estimated to affect 3 million people worldwide (prevalence of 1:2,000) and 200,000 people in the United States. The highest incidence is in Japan and the lowest in Israel, with both men and women affected equally. Like multiple sclerosis, narcolepsy also manifests itself in young adulthood with onset between the ages of 15 and 30 and eventually reaches a plateau with no progressive stage. Life expectancy is not altered. Interestingly, narcoleptic patients are more often born in March and less in September.

Narcolepsy is characterized by the common feature of excessive daytime sleepiness (EDS) in sudden bursts of 30 seconds and variations of cataplexy. Cataplexy is a sudden loss of muscle tone caused by strong emotions, such as laughter, anger, fear and excitement. Another feature of narcolepsy is abnormal rapid eye movement (REM) sleep such as sleep paralysis and hypnagogic hallucinations [1]. A normal sleep cycle progresses from wakefulness to the non-REM stage before residing in the REM stage. In contrast, a narcoleptic patient goes directly from full wakefulness to the REM cycle, completely bypassing the non-REM stage. The wakefulness state is characterized by full awareness, open sensory inputs and the ability to move muscles at will. During the non-REM state, one loses consciousness, sensory inputs are dampened and the muscles are atone. REM is a sleep state marked by limp muscles, an isolation from the environment, cerebral cortex reactivation, rapid bursts of eye movement and vivid dreaming by electroencephalography measurements.

Even though the term narcolepsy was first coined 130 years ago, research progress has been extremely slow with the initial association of this disease to the HLA gene region in the 1980s (Fig. 4.1). Both genetic and environmental factors appear to be responsible for this irreversible neuronal loss. The involvement of environmental factors was implicated from a study of a monozygotic twin pair positive for HLA DQB1*0602 in which only one showed symptoms of narcolepsy with cataplexy [2]. This is further corroborated with the prevalence of the disease which varies with geographic location and weather climate.

1877	First description in the medical literature
1880	Gelineau called the disorder "narcolepsy"
1902	Loewenfeld coined the term "cataplexy"
1935	First use of amphetamines in the treatment of narcolepsy
1960	Description of Sleep Onset REM periods in a narcoleptic subject
1970	Description of the Multiple Latency Test
1973	First report of a narcoleptic dog
1983	Association of narcolepsy with HLA-DR2
1985	Monoaminergic and cholinergic imbalance in narcolepsy
1992	Association of narcolepsy with HLA-DQB1*0602
1998	Identification of hypocretins/orexins and their receptors
1999	Hypocretin mutations cause narcolepsy in mice and dogs
2000	Human narcolepsy is also associated with an hypocretin deficiency

Figure 4.1 Research progress in narcolepsy over the past 100 years. [3]

Recent discoveries at the molecular level have implicated the involvement of the hypocretins/orexins and their receptors in narcolepsy. It was found that hypocretins decrease drowsiness and increase alertness, among other effects. Hypocretin-1 is a 33 amino acid molecule with a molecular weight of 3,562 Da whereas hypocretin-2 is a 28 amino acid linear peptide with a molecular weight of 2,937 Da. Studies in canine [4] and murine [5] models have shown that defective hypocretin transmission is responsible for the narcoleptic phenotype, mainly from the loss of hypocretin-1 and a decrease of hypocretin neurons in the gray matter. This observation,

along with its association with HLA gene susceptibility and the fact that no mutations or polymorphisms in hypocretin system genes were found, led to the theory that narcolepsy is an autoimmune disease. Subsequently, narcoleptic patients were shown to have low levels of hypocretin in the CSF [6, 7] and hypocretin neurons were selectively damaged [8, 9].

4.1.1.1 HLA associations

Canine narcolepsy was found to be caused by mutations in the gene encoding for hcrtr2, hypocretin receptor-2, in colonies of narcoleptic Dobermans and Labradors [4]. Even though the *Hcrtr2* gene is highly conserved between dog and human with a 97% similarity as shown in Figure 4.2, examinations revealed that mutations in hypocretin receptors were not the cause of human narcolepsy.

Instead, the genetic susceptibility of human narcolepsy is conferred by HLA DR2 and HLA DQ1 in linkage disequilibrium with HLA DQB1*0602 [10-12]. HLA DQB1*0602 is a specific marker for narcolepsy with cataplexy with 90% specificity [11]. However, most people positive for this susceptibility gene do not develop narcolepsy, conferring it only a specificity of 40% for narcolepsy in general [13]. This demonstrates that a genetic marker is not terribly useful as a molecular diagnostic.

A comparison of the amino acid sequence of HLA DQB1 in human to dog (DLA-DQB1) and mouse (H2-Ab1) shows a high percent similarity of 81% and 75%, respectively (Fig. 4.3). In fact, a majority of the amino acid substitutions in the narcolepsy susceptibility gene HLA DQB1*0602 are also present in dog and mouse. Since the HLA system contributes to T-cell mediated autoimmunity, perhaps aberration of this system in dogs and mice will allow narcoleptic animal models more similar to the cause of the human disease to be generated at a cheaper cost to facilitate molecular studies of narcolepsy development.

```
Hcrtr2 (dog)          1   MSGTKLEDSPPCRNWSSAPELNETQEPFLNPTDYDDEEFLRYLWREYLHPKEYEWVLIAG  60
Hcrtr2 (human)        1   MSGTKLEDSPPCRNWSSASELNETQEPFLNPTDYDDEEFLRYLWREYLHPKEYEWVLIAG  60
Hcrtr2 (mouse)        1   MSSTKLEDSLSRRNWSSASELNETQEPFLNPTDYDDEEFLRYLWREYLHPKEYEWVLIAG  60
Hcrtr2 (narc/Lab.)    1   MSGTKLEDSPPCRNWSSAPELNETQEPFLNPTDYDDEEFLRYLWREYLHPKEYEWVLIAG  60
Hcrtr2 (narc/Dob.)    1   MSGTKLEDSPPCRNWSSAPELNETQEPFLNPTDYDDEEFLRYLWREYLHPKEYEWVLIAG  60

Hcrtr2 (dog)          61  YIIVFVVALVGNVLVCVAVWKNHHMRTVTNYFIVNLSLADVLVTITCLPATLVVDITETW  120
Hcrtr2 (human)        61  YIIVFVVALIGNVLVCVAVWKNHHMRTVTNYFIVNLSLADVLVTITCLPATLVVDITETW  120
Hcrtr2 (mouse)        61  YIIVFVVALIGNVLVCVAVWKNHHMRTVTNYFIVNLSLADVLVTITCLPATLVVDITETW  120
Hcrtr2 (narc/Lab.)    61  YIIVFVVALVGNVLVCVAVWKNHHMRTVTNYFIVNLSLADVLVTITCLPATLVVDITETW  120
Hcrtr2 (narc/Dob.)    61  YIIVFVVALVGNVLVCVAVWKNHHMRTVTNYFIVNLSLADVLVTITCLPATLVVDITETW  120

Hcrtr2 (dog)          121 FFGQSLCKVIPYLQTVSVSVSVLTLSCIALDRWYAICHPLMFKSTAKRARNSIVIIWIVS  180
Hcrtr2 (human)        121 FFGQSLCKVIPYLQTVSVSVSVLTLSCIALDRWYAICHPLMFKSTAKRARNSIVIIWIVS  180
Hcrtr2 (mouse)        121 FFGQSLCKVIPYLQTVSVSVSVLTLSCIALDRWYAICHPLMFKSTAKRARNSIVVIWIVS  180
Hcrtr2 (narc/Lab.)    121 FFGQSLCKVIPYLQTVSVSVSVLTLSCIALDRWYAICHPLMFKSTAKRARNSIVIIWIVS  180
Hcrtr2 (narc/Dob.)    121 FFGQSLCKVIPYLQTVSVSVSVLTLSCIALDRWYAICHPLMFKSTAKRARNSIVIIWIVS  180

Hcrtr2 (dog)          181 CIIMIPQAIVMECSTMLPGLANKTTLFTVCDERWGGEIYPKMYHICFFLVTYMAPLCLMV  240
Hcrtr2 (human)        181 CIIMIPQAIVMECSTVFPGLANKTTLFTVCDERWGGEIYPKMYHICFFLVTYMAPLCLMV  240
Hcrtr2 (mouse)        181 CIIMIPQAIVMECSSMLPGLANKTTLFTVCDEHWGGEVYPKMYHICFFLVTYMAPLCLMI  240
Hcrtr2 (narc/Lab.)    181 CIIMIPQAIVMECSTMLPGLANKTTLFTVCDERWGGEIYPKMYHICFFLVTYMAPLCLMV  240
Hcrtr2 (narc/Dob.)    181 CIIMIPQAIVMECSTMLPGLANKTTLFTVCDERWGDPWNIICSSEKMEAPAACFTASRAR  240

Hcrtr2 (dog)          241 LAYLQIFRKLWCRQIPGTSSVVQRKWKPLQPASQPRGPGQQTKSRISAVAAEIKQIRARR  300
Hcrtr2 (human)        241 LAYLQIFRKLWCRQIPGTSSVVQRKWKPLQPVSQPRGPGQPTKSRMSAVAAEIKQIRARR  300
Hcrtr2 (mouse)        241 LAYLQIFRKLWCRQIPGTSSVVQRKWKQQQPVSQPRGSGQQSKARISAVAAEIKQIRARR  300
Hcrtr2 (narc/Lab.)    241 LAYLQIFRKLWCRQIPGTSSVVQRKWKPLQPASQPRGPGQQTKSRISAVAAEIKQIRARR  300
Hcrtr2 (narc/Dob.)    241 TADQVQD                                                  247

Hcrtr2 (dog)          301 KTARMLMVVLLVFAICYLPISILNVLKRVFGMFTHTEDRETVYAWFTFSHWLVYANSAAN  360
Hcrtr2 (human)        301 KTARMLMVVLLVFAICYLPISILNVLKRVFGMFAHTEDRETVYAWFTFSHWLVYANSAAN  360
Hcrtr2 (mouse)        301 KTARMLMVVLLVFAICYLPISILNVLKRVFGMFTHTEDRETVYAWFTFSHWLVYANSAAN  360
Hcrtr2 (narc/Lab.)    301 KTARMLMVVLLVFAICYLPISILNVLKRKV                            330

Hcrtr2 (dog)          361 PIIYNFLSGKFREEFKAAFSCCCLGVHHRQEDRLTRGRTSTESRKSLTTQISNFDNVSKL  420
Hcrtr2 (human)        361 PIIYNFLSGKFREEFKAAFSCCCLGVHHRQEDRLTRGRTSTESRKSLTTQISNFDNISKL  420
Hcrtr2 (mouse)        361 PIIYNFLSGKFREEFKAAFS-CCLGVHHRQGDRLARGRTSTESRKSLTTQISNFDNVSKL  419

Hcrtr2 (dog)          421 SEQVVLTSISTLPAANGAGPLQNW  444
Hcrtr2 (human)        421 SEQVVLTSISTLPAANGAGPLQNW  444
Hcrtr2 (mouse)        420 SEHVVLTSISTLPAANGAGPLQNW  443
```

Figure 4.2 **Comparison of amino acid sequences of Hcrtr2 between wild-type dog, human, mouse and narcoleptic dogs**. Amino acid residues from other sequences that differ from the wild-type dog are indicated in bold. Narcoleptic Labradors are represented as narc/Lab., narcoleptic Dobermans are represented as narc/Dob. [4]

```
HLA-DQB1 (human)       1   MSWKKALRIPGGLRVATVTLMLAMLSTPVAEGRDSPEDFVYQFKGMCYFTNGTERVRLVT    60
DLA-DQB1 (dog)         1   MSGKMTLCIPRGFWTAAVMMILVVLSIPVAEGRDSPQDFVYQFKFECYFTNGTERVRLLT    60
H2-Ab1   (mouse)       1   ----MALQIPSLLLSAAVVVLM-VLSSPGTEGGDSERHFVYQFMGECYFTNGTQRIRYVT    55
HLA-DQB1 (narc/human)  1   MSWKKALRIPGDLRVATVTLMLAMLSSLLAEGRDSPEDFVFQFKGMCYFTNGTERVRLVT    60

HLA-DQB1 (human)       61  RYIYNREEYARFDSDVGVYRAVTPLGPPDAEYWNSQKEVLERTRAELDTVCRHNYQ-LELR  120
DLA-DQB1 (dog)         61  KYIYNREEFVRFDSDVGEYRAVTELGRPDAEYWNPQKDEMDRVRAELDTVCRHNYG-VEEL  120
H2-Ab1   (mouse)       56  RYIYNREEYVRYDSDVGEHRAVTELGRPDAEYWNSQPEILERTRAELDTVCRHNYEGPETH  116
HLA-DQB1 (narc/human)  61  RYIYNREEYARFDSDVGVYRAVTPQGRPDAEYWNSQKEVLEGTRAELDTVCRHNYE-VAFR  120

HLA-DQB1 (human)       121 TTLQRRVEPTVTISPSRTEALNHHNLLVCSVTDFYPAQIKVRWFRNDQEETTGVVSTPLI  180
DLA-DQB1 (dog)         121 YTLQRRVEPTVTIFPSKTEVLNHHNLLVCSVTDFYPGQIKVRWFRNDQEQTAGVVSTPLI  180
H2-Ab1   (mouse)       117 TSLRRLEQPNVVISLSRTEALNHHNTLVCSVTDFYPTQIKVRWFRNGQEETVGVSSTQLI  176
HLA-DQB1 (narc/human)  121 GILQRRVEPTVTISPSRTEALNHHNLLVCSVTDFYPGQIKVRWFRNDQEETAGVVSTPLI  180

HLA-DQB1 (human)       181 RNGDWTFQILVMLEMTPQRGDVYTCHVEHPSLQNPIIVEWRAQSESAQSKMLSGIGGFVL  240
DLA-DQB1 (dog)         181 RNGDWTFQILVMLEMTPQRGDVYTCHVEHASLQSPITVQWRAQSESAQSKMLSGIGGFVL  240
H2-Ab1   (mouse)       177 RNGDWTFQVLVMLEMTPRRGEVYTCHVEHPSLKSPITVEWRAQSESAWSKMLSGIGGCVL  236
HLA-DQB1 (narc/human)  181 RNGDWTFQILVMLEMTPQRGDVYTCHVEHPSLQSPITVEWRAQSESAQSKMLSGVGGFVL  240

HLA-DQB1 (human)       241 GLIFLGLGLIIHHRSQKGLLH--------  261
DLA-DQB1 (dog)         241 GLIFLGLGLIIRHRSQKGLLH--------  261
H2-Ab1   (mouse)       237 GVIFLGLGLFIRHRSQKGPRGPPPAGLLQ  265
HLA-DQB1 (narc/human)  241 GLIFLGLGLIIRQRSQKGLLH--------  261
```

Figure 4.3 **Comparison of amino acid sequences of HLA DQB1 in human to dog (DLA-DQB1), mouse (H2-Ab1) and narcoleptic patients (HLA DQB1*0602)**. Amino acid residues from other sequences that differ from non-narcoleptic human are indicated in bold. Narcoleptic patients positive for the HLA DQB1*0602 susceptibility gene are represented as narc/human.

4.1.2 Current diagnostic tools for narcolepsy

Narcolepsy is either categorized as narcolepsy (i) with cataplexy, (ii) without cataplexy or (iii) due to medical condition. Common to all three is the presence of EDS. Current diagnosis is symptom-based and includes presence of cataplexy, a high Epworth Sleepiness Test score, a short sleep latency time, frequent sleep onset rapid eye movement periods (SOREMP) in the multiple sleep latency test (MSLT), and CSF hypocretin-1 level below 110 pg/ml. Even clinical diagnosis tools like the MSLT can result in false positives since ‘healthy’ individuals in a Wisconsin sleep cohort also displayed abnormal ‘narcolepsy-like’ MSLT [14]. These tools are used for

confirmation of the disease after symptoms have started appearing, which is long after the irreversible ablation of a critical number of hypocretin neurons [15, 16].

Early detection of narcolepsy is only possible with the discovery of novel molecular markers that appear either before or at the beginning of neuronal loss. These biomarkers will then serve a dual role for diagnosis and unraveling the molecular pathway to the pathogenesis of narcolepsy.

4.1.3 Proteomics studies on narcolepsy

To our knowledge, the study reported in this chapter is the first proteomics biomarker discovery in narcolepsy. It is aimed at the discovery of potential protein markers for narcolepsy with diagnostic utility and perhaps provide insight into the molecular basis of the disease. To achieve this, we analyzed the bound cargo of the carrier protein albumin in narcoleptic and control serum samples using the rapid readout technology described in Chapter 2.

4.1.4 Biomarker amplification

4.1.4.1 Hypothesis

The relative abundance of different proteins, in their intact, cleaved, or modified forms, is a consequence of the ongoing physiological and pathological events. Large proteins can only enter into circulation via active secretion or due to increased permeability of vascular walls as a result of a disease. On the other hand, cells and tissues in the body can respond to the current physiological state of the body by shedding protein fragments/peptides produced through endogenous enzymatic activities [21, 22] that permeate through the endothelial cell wall into blood circulation via passive diffusion. Peptide biomarkers already used in diagnostics include C-peptide for diabetes, amyloid Aβ 1-42 for Alzheimer's, and ACTH for adrenal insufficiency.

Unfortunately, due to the efficiency of the glomeruli in removing molecules smaller than 50kDa [23], any LMW molecules generated *in vivo* would be cleared quickly, reducing their concentration to undetectable levels. However, in the presence of excess high-abundance, high molecular weight (HMW) serum proteins, these low-abundance LMW molecules will bind to these carrier proteins (Fig. 4.4). This sequestration causes the LMW species to be protected from renal clearance and their circulation half-lives to be potentially prolonged, dependent on the half-life of the carrier protein For example, flavanoid and protoporphyrins are found to be 99% bound to albumin, and the immunoglobulin fragment, D3H44 Fab, increased its half-life from 0.8 hours to 10.4 hours when bound to albumin which has a half-life of 19 days, effectively increasing it's circulation concentration by 40 times [24]. A recent study showed that the high abundance polypeptides in plasma originate predominantly from major tissues such as liver, hematopoietic tissues and intestines with limited representation from the CNS, prostate and ovary [25]. Therefore, biomarkers specific for these underrepresented tissues will most likely be in low abundance and possibly enriched by carrier proteins, as evident in the case of ovarian cancer [26].

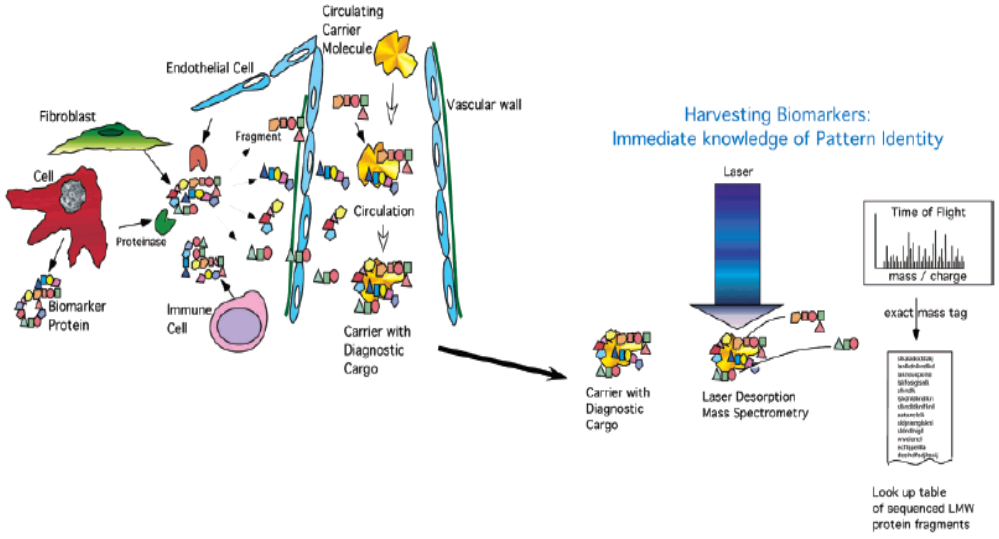


Figure 4.4 Biomarker amplification by carrier protein.[22]

In the current study, we focused on the carrier protein albumin. The advantages of investigating the albumin-bound subproteome is three-fold: (i) it acts like an *in vivo* fractionation step to simplify the proteome while simultaneously concentrate the low abundance potentially diagnostic fragments to the level that is detectable by MS [27, 28], (ii) numerous proteins/peptides with diagnostic potential bind to albumin [29] and (iii) these LMW molecules are well within the sensitivity and resolution mass region of 1,000 to 10,000 Da for MALDI TOF MS. These LMW peptides may be fragments of higher molecular weight inflammatory mediators such as cytokines and chemokines that may play a pivotal role in autoimmunity. Indeed, these, along with coagulation and complement factors, are well represented in the LMW serum proteome which has only been characterized recently [21, 30] (Fig. 4.5). Moreover, unlike many plasma proteins, human serum albumin is strictly regulated by colloidosmotic pressure feedback mechanisms, resulting in low inter- as well as intraindividual relative standard deviation below 7% [31]. Therefore, for the same amount of albumin captured, the variation in the bound species will allow differential diagnostic peptides to be uncovered.

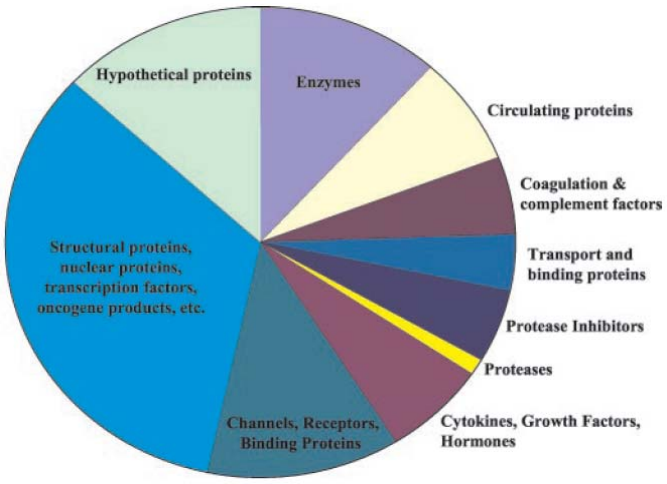


Figure 4.5 Pie chart representing the relative numbers of proteins identified within the LMW serum proteome. [21, 30]

4.1.4.2 Albuminome studies

LMW peptides and proteins that are selectively bound to albumin in human serum are termed the 'albuminome'. Common strategies that deplete high abundance proteins such as albumin have been shown to remove low abundance proteins such as diagnostic cytokines as well [17-20]. In recent years, many studies have demonstrated the diagnostic potential of these bound peptide or protein fragments [26, 32, 33], corroborating the biomarker amplification by carrier proteins hypothesis. The albumin-bound biomarker amplification concept has been proven by several studies. Lopez *et al*. [32] studied the albumin-bound LMW species and obtained a proteomic profiling fingerprint that is associated with Alzheimer's. Lowenthal *et al*. [26] studied the albuminome and discovered a subset of peptides that were specific for Stage I ovarian cancer, including fragments of low abundance molecules such as BRCA2 and tyrosine kinases. More recently, Lopez *et al.* [33] discovered several biomarker panels that differentiated Stage I ovarian cancer from patients without the disease.

4.2 MATERIALS AND METHODS

4.2.1 Study population and source of sera

A total of 30 clinical serum samples were obtained courtesy of Dr. Emmanuel Mignot from the Center for Narcolepsy at Stanford University. The samples were stratified into 3 groups based on the diagnosis condition and genetic composition: (1) patients diagnosed with narcolepsy and positive for the HLA DQB1*0602 marker gene, denoted NARC/+ (n=10), (2) patients not diagnosed with narcolepsy but positive for the HLA DQB1*0602 marker gene, denoted CTRL/+ (n=10) and (3) patients not diagnosed with narcolepsy and negative for the HLA DQB1*0602 marker gene, denoted CTRL/- (n=10).

4.2.2 Sample preparation

All samples were stored at -80°C before analysis. Each serum sample was diluted 1:10 in triplicate and albumin was captured selectively, followed by the elution of the bound peptide and protein fragments into a clean collection plate using the ProXPRESSION Biomarker HT Enrichment Kit (PerkinElmer). The technology is essentially an affinity chromatography with Cibachron blue dye which binds to albumin. The proprietary matrix in the kit where the Cibachron blue is immobilized retains the captured albumin carrier protein, allowing the subsequent elution step to only disrupt the binding of the peptide and protein fragments to albumin and releases them for analysis. This was performed using a PerkinElmer MultiPROBE II PLUS HT EX liquid handler. The bound cargo in 96-well microtiter plate was subsequently concentrated by vacuum centrifugation and resuspended in 1X PBS, pH 7.4 prior to incubation on IMAC30 ProteinChip arrays (Bio-Rad).

All ProteinChip arrays were processed on the same day in a 96-well format as follows: IMAC30 ProteinChip arrays from the same batch were activated with 50 mM nickel (II) sulfate

hexahydrate three times for 15 mins with gentle shaking, followed by a quick rinse with HPLC grade water. The arrays were then equilibrated with 100 mL 1X PBS on a shaker at room temperature for 15 mins, thrice. The buffer was discarded and remaining droplets were aspirated from the spots using a vacuum tip. The IMAC30 arrays were then loaded into a ProteinChip bioprocessor cassette to facilitate high-throughput analysis. The serum albumin-derived samples were deposited onto the array spots using a liquid handler. Each serum sample was run in triplicate. The bioprocessor cassette was sealed with aluminum foil, placed in a humidifying chamber, stored in a vacuum desiccator and incubated overnight in 4°C. The following day, the serum samples were removed and the arrays were washed with 200 µL 1X PBS three times for 15 mins with gentle shaking at room temperature. The bioprocessor was subsequently removed and any remaining droplets were aspirated from the spots using a vacuum tip. The ProteinChips were allowed to air dry for 15 mins. Once dry, two 1 µL aliquots of 5 mg/mL α-cyano-4-hydroxycinnamic acid (CHCA) matrix (LaserBio Labs, France) were added to each spot. All washing steps and matrix deposition were performed using a liquid handler. Matrix solution was kept protected from light at room temperature until ready for use. The spots were allowed to air dry before prOTOF MALDI-TOF mass spectrometry analysis.

Analytical variables optimization for maximal mass peak production was performed as described above using a standard serum sample from the NIST on IMAC30 ProteinChip arrays. For protein identification, serum albumin-bound cargo was obtained as described above and loaded onto IMAC spin columns, pre-charged with nickel. Bound species were eluted and subject to FT-MS analysis.

4.2.3 MALDI TOF mass spectrometry analysis

ProteinChip arrays were placed in a custom made adapter for mass spectrometry analysis in the prOTOF2000 MALDI O-TOF mass spectrometer interfaced with TOFWorks software (PerkinElmer/SCIEX). Its orthogonal design enabled a single external mass calibrant to achieve better than 5 ppm mass accuracy over the 1,000 to 10,000 mass range acquired. A 2-point external calibration of the prOTOF instrument was performed before acquiring the spectra. The batch analysis of the 12 chips was split into two runs of six array chips each. This is constrained by the plate holder which can only accommodate a maximum of six chips. The sixth chip per run has three spots dedicated for the NIST reference serum sample (positive control), buffer blank (negative control) and calibrant (instrument calibration) to ensure the integrity of the whole process. Acquisition was performed with a laser intensity of 65% at 100 Hz, 30V declustering voltage, 150 mL/min cooling flow rate, and 200 mL/min focusing flow rate. Mass spectra were acquired in a circular pattern. The prOTOF data files generated an average of 1 million data points per spectrum.

The same ProteinChip arrays were removed from prOTOF and loaded into the Ciphergen PBS-IIc mass spectrometer. Instrument calibration was performed externally with the All-in-1 Peptide Calibrant (Ciphergen). Acquisition was performed in a batch mode of 12 arrays with a laser intensity of 170 and a sensitivity of 9. Mass spectra were acquired in a linear pattern on the same spots. The PBS-IIc data files generated an average of 40,000 data points per spectrum.

4.2.4 Biostatistical analysis

Raw spectra from the prOTOF were exported as text files using the prOTOF loader program and preprocessed to restore the zero-intensity values [32, 34]. The 30 serum samples run in triplicate generated a total of 90 high-resolution mass spectra. Spectra from two serum samples with

macroscopic blood contamination were excluded from the analysis. The total ion current (TIC) of each spectrum was calculated and the average TIC was computed across the remaining 84 spectra. Spectra with a TIC value that was greater than twice or less than half of the average TIC were deemed outliers and were omitted from the study. Global normalization of the signal intensity of the mass peaks was performed by normalizing to the average TIC of the remaining 68 spectra.

All spectra were run through the Progenesis PG600 software (Nonlinear Dynamics, UK) for peak detection using the following parameters to remove background noise: noise filter size 4, background filter size 70, and isotope detection in MALDI mode with peak threshold 25 and window 0.1 Da.

The same mass spectral data set was analyzed by four distinct algorithms. This serves to uncover consensus, differential mass peaks that are less prone to biases to a particular algorithm, as previously described in Section 2.4.2. All but the t-test analysis were performed by the same operator.

Analyses were performed to discover statistically differential markers between the following groups: NARC/+ vs CTRL/±, NARC/+ vs CTRL/+, CTRL/+ vs CTRL/-, and GENE/+ vs GENE/-. ROC curve analysis was then performed using SAS on the differential peaks to determine their discriminatory power.

4.2.5 Protein identification

Samples were incubated in a denaturing solution of 8 M urea/1% SDS/100 mM ammonium bicarbonate/10 mM DTT pH 8.5 at 37ºC for 1 h. Next, the samples were alkylated for 1 h by the addition of iodoacetamide to a final concentration of 40 mM and then quenched with 2 M DTT. Following the addition of 4X LDS loading buffer (Invitrogen), each sample was centrifuged at 14,000 rpm for 5 mins at room temperature, and each sample was fractionated on a NuPAGE

10% Bis-Tris 10 lane gel (Invitrogen) for 2.5 h at 125 volts, 50 mA and 8W. Gels were shrunk overnight by the addition of 50% ethanol and 7% acetic acid, and then allowed to swell for 1 h by the addition of deionized water. Gels were stained with SimplyBlue Safe Stain (Invitrogen) for 2-4 h, imaged, and sliced horizontally into fragments of equal size based on the molecular weight markers.

In-gel digestion was performed after destaining and rinsing the gel sections with two washes of 50% ethanol and 7% acetic acid, followed by two alternating washes with 50 mM ammonium bicarbonate and acetonitrile. After removal of the last acetonitrile wash, 100 µL of sequencing grade trypsin (Promega) was added to each gel slice at a concentration of 6.6 µg/ml in 50 mM ammonium bicarbonate/10% acetonitrile. The gel slices were allowed to swell for 30 mins on ice, after which the tubes were incubated at 37ºC for 24 h. Peptides were extracted with one wash of 100 µL of 50 mM ammonium bicarbonate/10% acetonitrile and one wash of 100 µL of 50% acetonitrile/0.1% formic acid. The extracts were pooled and frozen at -80ºC, lyophilized to dryness and redissolved in 40 µL of 5% acetonitrile, 0.1% formic acid.

Samples were then loaded into a 96-well plate (AbGene) for mass spectrometry analysis on both a Thermo Fisher Scientific LTQ-FT and Thermo Fisher Scientific LCQ Deca XP Plus. For each run, 10 µL of each reconstituted sample was injected with a Famos Autosampler, and the separation was performed on a 75 µm x 20 cm column packed with C_{18} Magic media (Michrom Biosciences) running at 250 nl/min provided from a Surveyor MS pump with a flow splitter with a gradient of 5-60% water 0.1% formic acid, acetonitrile 0.1% formic acid over the course of 120 mins (150 mins total run). Between each set of samples, standards from a mixture of 5 angiotensin peptides (Michrom Biosciences) were run for 2.5 h to ascertain column performance and observe any potential carryover that might have occurred. The LTQ-FT was run in a top nine configuration with one MS 200K resolution full scan and nine MS/MS scans and the LCQ Deca XP Plus was run in a top five configuration with one MS full scan and five MS/MS

scans. Dynamic exclusion was set to 1 with a limit of 180 s with early expiration set to 2 full scans.

SEQUEST-identified peptide sequences, protein accession numbers, and MALDI TOF MS *m/z* values corresponding to the differential peaks were entered into the FRAGMINT software to generate candidate protein fragments consistent with the identified protein and peptide sequences and the observed MALDI TOF MS *m/z* values as described [35].

4.2.6 Western blot analysis

Serum samples from each group were separated by 10% Tris-Glycine-SDS-polyacrylamide gel electrophoresis and transferred to an Immobilon-PSQ 0.2 μm PVDF membrane (Millipore). Western blotting was performed on 50 μg total protein per lane. The membrane was blocked with 5% milk in 1X TBST overnight and then probed for 1 h with a goat polyclonal antibody to Bikunin (sc-21597, Santa Cruz Biotechnology Inc.) diluted 1:2,000. The membrane was washed with 1X TBST twice for 15 mins and then with 1X PBS twice for 15 mins, followed by incubation for 1 h with HRP-labeled rabbit anti-goat secondary antibody (Bio-Rad) diluted 1:2,000. The washing steps were repeated before detection with SuperSignal West Pico Chemiluminescent Substrate (Thermo Scientific). The membrane was stripped, blocked and re-probed with rabbit polyclonal antibody against transthyretin (sc-13098, Santa Cruz Biotechnology Inc.) diluted 1:2,000 and HRP-labeled goat anti-rabbit secondary antibody (Bio-Rad) diluted 1:2,000. All antibodies were diluted in 5% milk in 1X TBST. All incubations were performed at room temperature. ImageJ was used for quantitation.

4.3 RESULTS

4.3.1 Analytical variables assessment

4.3.1.1 Evaluation of albumin-bound subproteome

In our study, we intend to simplify the albumin-bound subproteome one step further via an additional fractionation step by only analyzing species that bind to a metal affinity capture surface (IMAC30). As this will reduce the amount of analytes presented to the mass spectrometer during data acquisition, we investigated the dilution factor of serum that would optimize the binding of albumin to the enrichment platform, and indirectly optimize the recovery of the bound species. A standard serum sample obtained from the NIST was either diluted 10-fold or 20-fold prior to loading onto the albumin enrichment plate. Bound analytes were subsequently eluted per manufacturer's protocol. The eluate was introduced to an IMAC30 surface charged with nickel in its entirety and analyzed by mass spectrometer (Fig. 4.6). It was apparent that a 10-fold diluted loading maximized the output mass peaks (Panels A and B, Fig. 4.6). A lower dilution factor was not investigated to conserve precious clinical samples that will be run in triplicate in the profiling study for validation purposes.

For comparison, the native serum sample along with the bound cargo and flow-through fractions from the albumin enrichment step were run on an IMAC30 ProteinChip and analyzed. Figure 4.7 shows that the cargo fraction produced a similar profile to that of the native sample with increased intensity and species detection for the LMW population. This substantiates the argument that pre-fractionation is an integral part of studies involving complex biological samples as it reduces the effect of ion suppression considerably.

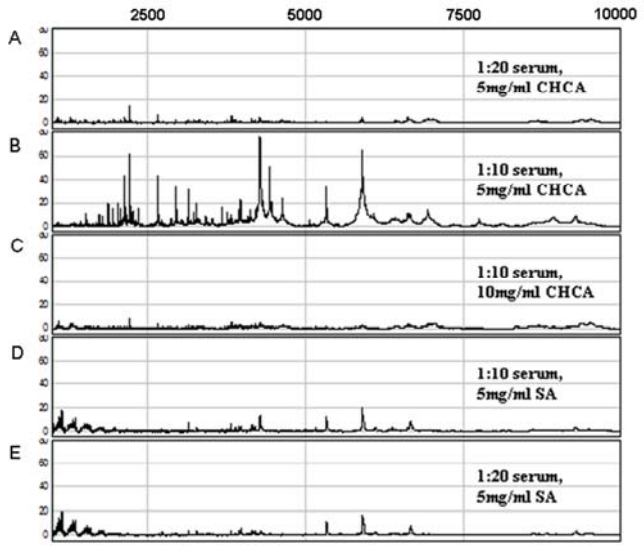


Figure 4.6 **Optimization of serum dilution factor and ionizing matrix concentration**. Shown are mass spectra corresponding to the same serum sample diluted 10- and 20-fold and evaluated with either CHCA (A, B) or SA (D, E). CHCA was also evaluated at different concentrations, 5 mg/ml (A, B) and 10 mg/ml (C). Signal intensity (y-axis) is plotted against the mass range, *m/z* (x-axis).

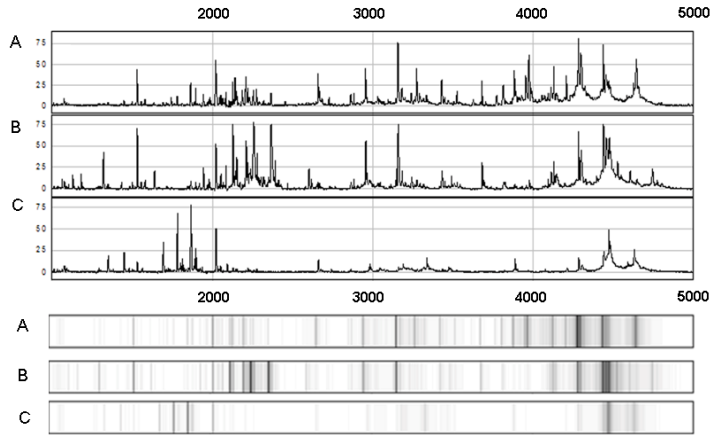


Figure 4.7 **Comparison of native serum sample to the albumin-enriched and albumin-depleted fractions**. Shown are the mass spectra of the same serum sample analyzed in its native (A), albumin-enriched (B) and albumin-depleted (C) forms. The corresponding gel-views are displayed below the mass spectra.

4.3.1.2 Surface retentate chemistry and matrix choice

IMAC30 charged with nickel ions optimized in Section 2.3.2 to generate the most number of peaks was used in this study.

The two most common matrices used in MALDI TOF MS analysis, CHCA and SA, were evaluated at the 5 mg/ml and 10 mg/ml concentrations. CHCA conferred more peaks for this particular subset of the serum proteome than SA at the preferred concentration of 5 mg/ml (Fig. 4.6).

4.3.1.3 Spectral reproducibility

The reproducibility of the overall protein profile on the ProteinChip-MS platform was also evaluated. Three serum samples (one from each group being profiled) were run in duplicate on the same IMAC30 ProteinChip. The high accuracy and resolution conferred by the prOTOF mass spectrometer as discussed in Section 2.3.3 enabled reproducible replicate spectra from individual samples to be obtained (Fig. 4.8).

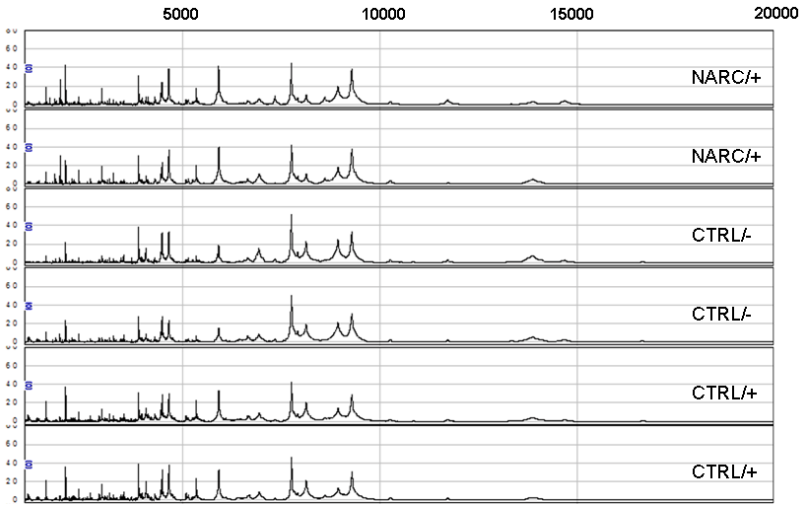


Figure 4.8 **Spectral reproducibility was evaluated with a sample from each group run in duplicate on IMAC30 chip surface**. NARC/+ represents serum sample from narcolepsy patients positive for the HLA DQB1*0602 susceptibility gene, CTRL/+ represents serum sample from non-narcolepsy patients positive for the HLA DQB1*0602 susceptibility gene, CTRL/- represents serum sample from non-narcolepsy patients negative for the HLA DQB1*0602 susceptibility gene.

4.3.2 Profiling of narcolepsy and non-narcolepsy sera

The 30 serum samples were run in triplicate to minimize analytical variance in the technical process. The workflow employed in this study is shown in Figure 4.9.

In this study, four distinct classification algorithms were applied on the same narcolepsy mass spectral data set as described in Section 2.4.2. Four comparisons were performed between the three sample groups. The first comparison seeks biomarkers for narcolepsy in general, without taking genetic predisposition into consideration (NARC/+ versus CTRL/±). The second comparison seeks markers that are specific for narcolepsy with HLA DQB1*0602 genetic

predisposition (NARC/+ versus CTRL/+). The third comparison seeks markers for HLA DQB1*0602 positivity without developing narcolepsy (CTRL/+ versus CTRL/-). The final comparison seeks biomarkers specific for the HLA DQB1*0602 gene positivity (GENE/+ versus GENE/-, where GENE/+ = NARC/+ and CTRL/+, and GENE/- = CTRL/-). Although our interest lies in narcolepsy-specific markers, the latter two comparisons are reported for completion.

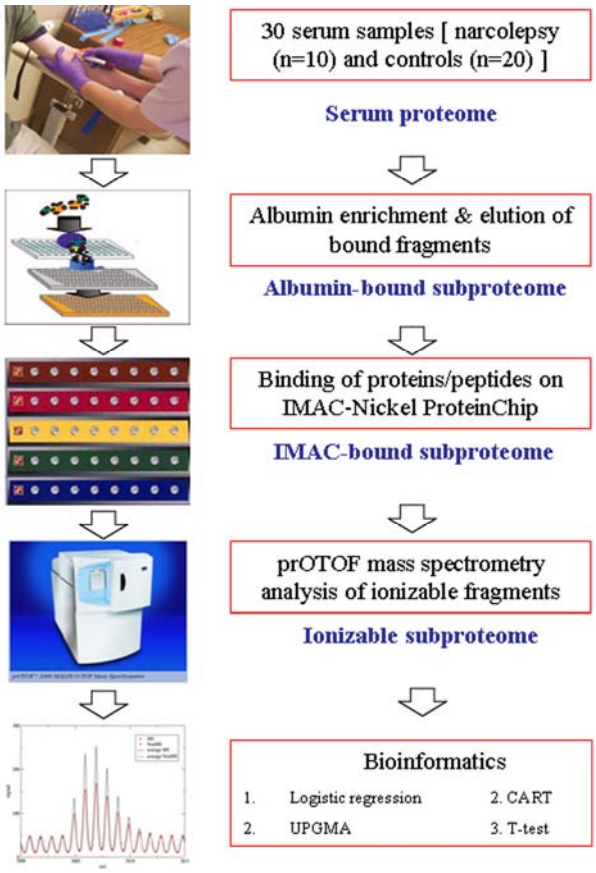


Figure 4.9 Blood-based biomarker discovery workflow.

Table 4.1 shows the list of candidate markers for each comparison aforementioned. All of the listed mass peaks appeared as potential markers with an expression difference that is statistically significant in at least two of the four bioinformatics platforms employed. These consensus peaks reduce the possibility of investigating mass peaks that are a result of overfitting

to a specific algorithm and confer a higher confidence in their true discriminatory ability. ROC analysis was performed on all four models and their respective diagnostic accuracy measures are tabulated in Table 4.2.

Comparison	Mass peaks (*m/z*)
I. NARC/+ vs. CTRL/±	1740.94
	1809.98
	3826.00
	5077.83
II. NARC/+ vs. CTRL/+	1896.06
	2036.08
	3826.00
III. CTRL/+ vs. CTRL/-	2080.98
	2127.02
	2209.09
	4296.17
IV. GENE/+ vs. GENE/-	1938.09
	1943.91
	2127.01
	5077.83

Table 4.1 **Differential mass peaks discovered from logistic regression, CART, UPGMA hierarchical clustering and *t*-test**. The consensus peaks listed appeared as statistically differential across at least two of the four classification algorithms applied to the same narcolepsy mass spectral data set. CTRL/± represents serum samples from all non-narcolepsy patients, GENE/+ represents serum samples from patients positive for the HLA DQB1*0602 susceptibility gene, GENE/- represents serum samples from patients negative for the HLA DQB1*0602 susceptibility gene.

ROC analysis of the differential peaks for each group comparison

Comparison	Sensitivity (%)	Specificity (%)	PPV (%)	AUC	Percent Accuracy (%)
I. NARC/+ vs. CTRL/±	63.16	82.22	85.96	0.79	76.56
II. NARC/+ vs. CTRL/+	68.40	83.33	90.32	0.83	75.67
III. CTRL/+ vs. CTRL/-	72.22	66.67	77.50	0.71	68.89
IV. GENE/+ vs. GENE/-	81.08	44.44	73.68	0.66	65.63

Table 4.2 **Diagnostic accuracy measures for each group comparison**. ROC analysis was performed on models consisting of differential mass peaks from Table 4.1.

In all comparisons, all potential biomarkers are molecules found in differential abundance between the two comparison groups. Of interest to us are the differential mass peaks that are responsible for narcolepsy onset in general (comparison I) or responsible for narcolepsy onset in genetically susceptible people (comparison II). These two models have comparable diagnostic potential even though unique peaks were selected as statistically differential in their respective disease classification model. The protein profiles obtained in this study to discriminate between narcolepsy and controls displayed good discriminatory ability with an AUC of 0.80. Models from these two comparisons on the average presented a PPV of 88%, a high specificity at 83% and a decent sensitivity at 66%. Differential peaks in comparison III would encompass protein/peptides that are protective against narcolepsy in the wake of genetic susceptibility whereas peaks in comparison IV would represent molecules that are a product of the HLA DQB1*0602 gene.

4.3.3 Biomarker identification

The platform that was employed to discover the differential peaks above is not capable of providing fragmentation data for identification. Therefore, the ProteinChip surface chemistry was recapitulated on IMAC spin columns and the enrichment protocol was scaled up to obtain a sample suitable for analysis by tandem Fourier transform-mass spectrometry (FT-MS). Since all the differential peaks of interest as listed in Table 4.1 were LMW peptides, the enriched sample was first separated by SDS-PAGE. The less than 15 kDa region was then excised, reduced, alkylated, digested and analyzed by tandem FT-MS. On the basis of the SEQUEST-identified peptides, the FRAGMINT software was used to generate candidate fragment identity for three differential peaks which were consistent with their discriminatory *m/z* values. Two of the identified fragments were from the GENE/+ versus GENE/- comparison, and one from the NARC/+ versus CTRL/+ comparison (Table 4.3). As we are more interested in the identity of candidate biomarkers for differentiating between narcolepsy and control, verification of the unique putative marker from the latter comparison was the prime focus. The peak at *m/z* = 2036 was assigned to the peptide sequence RGPCRAFIQLWAFDAVK, which corresponds to the bikunin sequence of the α-1-microglobulin/bikunin precursor (AMBP). This mass peak was higher in the NARC/+ samples over the CTRL/+ (Fig. 4.10).

Group Comparison	Mass Peak (*m/z*)	Higher Expression in	FRAGMINT Identifications
GENE/+ vs. GENE/-	1938.09	GENE/-	Serum Amyloid A4
GENE/+ vs. GENE/-	1943.91	GENE/-	Albumin
NARC/+ vs. CTRL/+	2036.08	NARC/+	α1-microglobulin/bikunin

Table 4.3 **Candidate identifications from group comparisons**. Identifications were obtained from FT-MS SEQUEST-identified peptides and FRAGMINT.

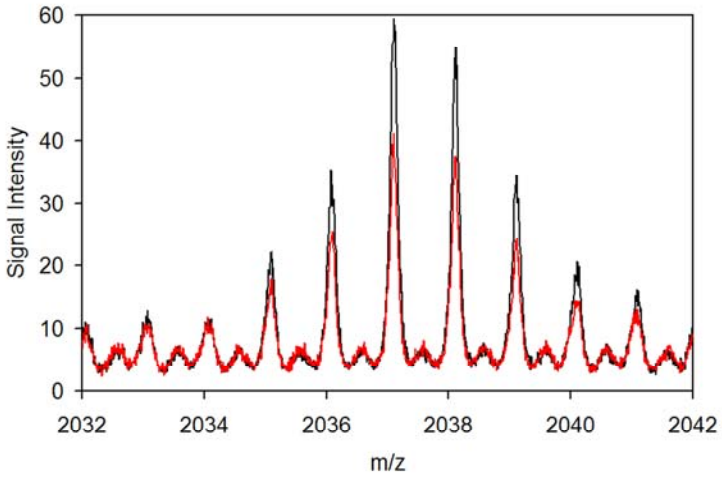
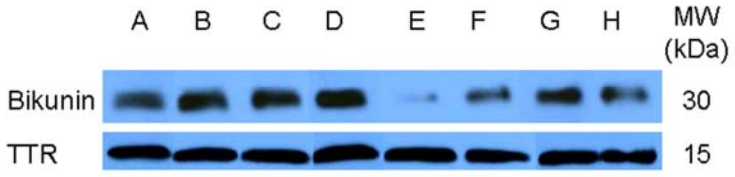


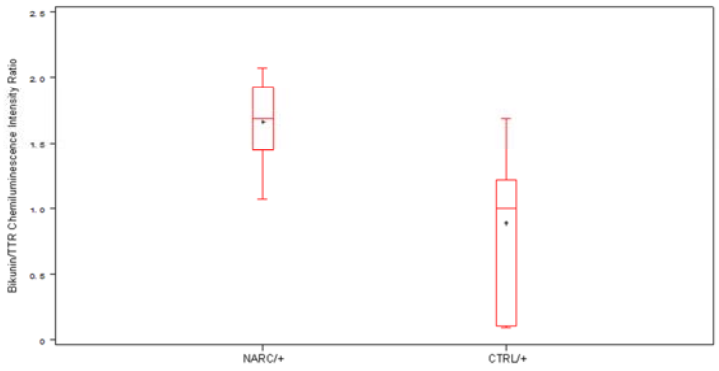
Figure 4.10 **Differential marker for narcolepsy.** Overlay view of mean signal intensity of peak 2036 in NARC/+ to CTRL/+ group comparison. Black trace, narcolepsy; red trace, control.

4.3.4 Validation of bikunin as a potential biomarker for narcolepsy

We assessed the level of bikunin in both the NARC/+ and CTRL/+ groups in serum by Western blot analysis. Bikunin protein levels were ascertained with an antibody that recognizes an epitope within the bikunin sequence identified. 50 µg of total serum protein was loaded for each sample on a 10% gel. Transthyretin (TTR) was probed as a loading control for serum. All samples from these two groups that were not omitted from the study due to hemoglobin contamination and whose mass spectra were not omitted as outliers (8 from NARC/+ and 7 from CTRL/+) were run and probed for bikunin concurrently to minimize technical bias. It was found that the level of bikunin was elevated in the NARC/+ serum samples over CTRL/+ (Panel A, Fig. 4.11). The chemiluminescence intensity was quantified and the ratio of bikunin intensity over TTR intensity was calculated. The bikunin/TTR ratio between the NARC/+ and CTRL/+ groups showed a statistically significant difference with a p-value of < 0.05. The box and whisker plot of the ratios are shown in Figure 4.11, Panel B. The protein was more abundant in the serum of NARC/+ patients compared with CTRL/+ cases. Based on this correlation between the Western blot data for bikunin protein and the mass spectrometry data, we postulate that peak at $m/z = 2036$ might represent a degraded fragment of bikunin, a protein which could play a role in the development of narcolepsy.



A



B

Figure 4.11 **Differential levels of bikunin between narcolepsy and control groups**. (A) Representative Western blot depicting levels of bikunin in the serum of narcolepsy and control patients. A-D= NARC/+ sera, E-H=CTRL/+ sera. TTR was used as loading control. (B) Box and whisker plot of Bikunin/TTR chemiluminescence intensity ratio.

4.4 DISCUSSION

A common practice in biomarker discovery is to compare controls to patients with an established disease with the assumption that markers specific for the disease may be in higher abundance in the latter group. However, if the disease markers to be discovered are aimed at early detection, patients who are susceptible to develop the disease will constitute a more appropriate group of samples as these 'pre-disease' patients may produce biomarkers that are different from those of established patients. In addition, investigations of an uncommon disease like narcolepsy calls for case-control studies, which confer a significant advantage especially when hypotheses include gene-environment interactions. The search for environmental risk factors is facilitated if cases

and controls are chosen to be similar on this genetic marker. Therefore, in our study, we have enlisted as controls patients who are positive for the HLA DQB1*0602 gene and hence susceptible to narcolepsy, in the hope of discovering markers that are applicable to the early detection of narcolepsy.

Figure 4.7 showed that the protein profiles of the native serum sample and the albumin-bound fraction were very similar, suggesting most of the peaks observed in serum profiling studies might have originated predominantly from albumin-bound species, as observed by others [36]. Interestingly, the flow-through fraction revealed peaks that were either not detected or were of extremely low intensity in the native sample in the less than 2,000 *m/z* region. Even though it might be interesting to analyze the flow-through fraction for novel, low abundance biomarkers, this was not undertaken in the present study as the focus was in carrier protein-bound cargo and simply because the composition of each flow-through fraction cannot be standardized across samples which could introduce bias.

In this study, we generated and analyzed 90 serum albumin-derived proteome profiles from 30 patients to discover biomarkers specific for narcolepsy. The potential of using proteomic differential protein pattern profiling can be evaluated as follows. In the case where a patient who is positive for the HLA DQB1*0602 susceptibility gene is being evaluated for the risk of developing narcolepsy, specificity is of importance over sensitivity. In this scenario, the current model from the NARC/+ versus CTRL/+ comparison could be used as a confirmatory test with a specificity of 83% and a PPV of 90%. However, sensitivity takes priority if the goal is for screening. In this case, the model obtained from the NARC/+ versus CTRL/± comparison will not be a good platform to detect those prone to develop narcolepsy since it suffers from a low sensitivity of only 63%. However, its specificity of 82% is still much greater than that of the genetic marker HLA DQB1*0602 [13].

The advantage of analyzing samples from patients with the same genetic background is that it will allow us to attribute potential biomarkers found to environmental factors that

propagate disease susceptibility to realization. Identification of these candidate biomarkers is thus crucial to the understanding of the molecular mechanisms underlying the pathogenesis of narcolepsy.

A peak at *m/z* = 2036 was found to have a preferential presence in narcolepsy over control samples with the same genetic background. This peak was subsequently identified as part of the bikunin protein. To evaluate its potential as a disease marker, serum samples from both NARC/+ and CTRL/+ groups were probed with an antibody whose epitope is within the identified bikunin sequence. The relative level of the bikunin protein was found to correlate well to that of the 2036 peak in our proteomic analysis, with a higher level in the narcolepsy group. Given these data, we believe that the molecule identified in our study with *m/z* = 2036 is a fragment of bikunin, specifically from Kunitz domain II which confers anti-inflammatory property. We speculate that the degraded bikunin fragment was prevented from renal clearance and by extension enriched through its binding to albumin, as in the case of α1-microglobulin (the other unrelated protein product from the common AMBP precursor) found in human plasma to be complexed to IgA and albumin [37]. Our results support, but do not prove, the causal relationship between bikunin and narcolepsy onset. This remains to be fully validated in a larger study as this pilot study is admittedly limited in sample size. A *post hoc* analysis for the difference observed in the Western blots suggested that at least eight samples per group being compared will have to be represented in order to achieve a desired power of 90%. Nonetheless, it would not be surprising if bikunin assumes a role via its inflammatory property in narcolepsy development.

A growing body of evidence has indicated that bikunin is involved in many pathophysiological processes, such as immune response, inflammation, tumorigenesis, and metastasis [38, 39]. Bikunin plays a role in inflammation and innate immunity because of its tandem Kunitz-type binding domains, which has anti-proteolytic and anti-inflammatory properties. Free bikunin molecules are present in plasma at an insignificant level. More than 98% of bikunin present in circulation is complexed to the proinhibitors, inter-α-inhibitor and pre-α-

inhibitor, where bikunin remains inactive until its release through degradation by elastase at sites of inflammation [39]. Free or cell-bound bikunin are predicted to downregulate cytokine expression, render macrophages/neutrophils less active and impair inflammatory processes [40].

Increased free, uncomplexed bikunin in blood and urine has been shown to correlate well with inflammatory conditions [41, 42] and to be better predictors of vascular inflammation than existing biomarkers [43]. Higher level of bikunin has also been reported in chronic inflammations in autoimmune disorders such as RA and systemic lupus erythematosus [43]. Narcolepsy is generally believed to be an autoimmune disease that results in the irreversible loss of hypocretin-producing neurons. Although evidence to support the autoimmunity cause of narcolepsy has remained elusive, it is not an unlikely possibility as the detection of bikunin mRNA has been shown in rat neurons [44, 45] and human astroglias of Alzheimer's patients in brain regions where loss of neurons was observed [46, 47]. Long-standing inflammation leading to neuronal loss may be present before clinical symptoms are presented.

Our primary goal was to evaluate the MS-based platform described here as a diagnostic and discovery tool in the hunt for biomarkers specific for uncommon diseases such as narcolepsy. We have shown that these differential marker peaks can serve as reliable candidates for downstream validation efforts and confirmed the identification of bikunin as one of the differential peaks by Western blot. It is also foreseeable that this methodology has clinical utility as a complement to existing diagnostic tools to analyze samples from patients who have first been screened for the HLA DQB1*0602 susceptibility gene. HLA genotyping is useful in prescreening before other tests because susceptible individuals can be identified early and genetic susceptibility does not change with time or age. Therefore, the samples used in this narcolepsy study are representative of the samples that will undergo the same processing procedure for diagnosis in the clinical setting. The preliminary results reported here suggest its potential as a non-invasive diagnostic tool as demonstrated by the discriminatory protein profiles obtained with higher

specificity than the genetic marker HLA DQB1*0602. This potential will likely be realized pending efforts to further validate the robustness and reproducibility of this platform.

4.5 CONCLUSION

Serum represents an ideal biological sample as it is rich in proteins and can be obtained non-invasively. Our preliminary validation of bikunin as a potential player in the pathogenesis of narcolepsy demonstrates the applicability of this platform in biomarker discovery. Confirmation of bikunin's role in a larger validation set or the discovery of other novel biomarkers using this methodology will no doubt fill the void created by the lack of molecular markers for narcolepsy. Subsequent therapeutics development targeting these markers will provide an avenue for controlling the disease instead of the symptoms. Even though narcolepsy is non-life threatening, a delay in its onset will significantly improve the quality of life of those susceptible to it.

4.6 BIBLIOGRAPHY

1. Yoss RE, Daly DD: **Criteria for the diagnosis of the narcoleptic syndrome**. *Proc Staff Meet Mayo Clin* 1957, **32**(12):320-328.
2. Dauvilliers Y, Maret S, Bassetti C, Carlander B, Billiard M, Touchon J, Tafti M: **A monozygotic twin pair discordant for narcolepsy and CSF hypocretin-1**. *Neurology* 2004, **62**(11):2137-2138.
3. Mignot E: **A hundred years of narcolepsy research**. *Archives italiennes de biologie* 2001, **139**(3):207-220.
4. Lin L, Faraco J, Li R, Kadotani H, Rogers W, Lin X, Qiu X, de Jong PJ, Nishino S, Mignot E: **The sleep disorder canine narcolepsy is caused by a mutation in the hypocretin (orexin) receptor 2 gene**. *Cell* 1999, **98**(3):365-376.
5. Chemelli RM, Willie JT, Sinton CM, Elmquist JK, Scammell T, Lee C, Richardson JA, Williams SC, Xiong Y, Kisanuki Y *et al*: **Narcolepsy in orexin knockout mice: molecular genetics of sleep regulation**. *Cell* 1999, **98**(4):437-451.
6. Nishino S, Ripley B, Overeem S, Nevsimalova S, Lammers GJ, Vankova J, Okun M, Rogers W, Brooks S, Mignot E: **Low cerebrospinal fluid hypocretin (Orexin) and altered energy homeostasis in human narcolepsy**. *Ann Neurol* 2001, **50**(3):381-388.
7. Dauvilliers Y, Baumann CR, Carlander B, Bischof M, Blatter T, Lecendreux M, Maly F, Besset A, Touchon J, Billiard M *et al*: **CSF hypocretin-1 levels in narcolepsy, Kleine-Levin syndrome, and other hypersomnias and neurological conditions**. *J Neurol Neurosurg Psychiatry* 2003, **74**(12):1667-1673.
8. Blouin AM, Thannickal TC, Worley PF, Baraban JM, Reti IM, Siegel JM: **Narp immunostaining of human hypocretin (orexin) neurons: loss in narcolepsy**. *Neurology* 2005, **65**(8):1189-1192.

9. Thannickal TC, Moore RY, Nienhuis R, Ramanathan L, Gulyani S, Aldrich M, Cornford M, Siegel JM: **Reduced number of hypocretin neurons in human narcolepsy**. *Neuron* 2000, **27**(3):469-474.

10. Billiard M, Seignalet J, Besset A, Cadilhac J: **HLA-DR2 and narcolepsy**. *Sleep* 1986, **9**(1 Pt 2):149-152.

11. Mignot E, Hayduk R, Black J, Grumet FC, Guilleminault C: **HLA DQB1*0602 is associated with cataplexy in 509 narcoleptic patients**. *Sleep* 1997, **20**(11):1012-1020.

12. Mignot E, Lin X, Arrigoni J, Macaubas C, Olive F, Hallmayer J, Underhill P, Guilleminault C, Dement WC, Grumet FC: **DQB1*0602 and DQA1*0102 (DQ1) are better markers than DR2 for narcolepsy in Caucasian and black Americans**. *Sleep* 1994, **17**(8 Suppl):S60-67.

13. Dauvilliers Y, Arnulf I, Mignot E: **Narcolepsy with cataplexy**. *Lancet* 2007, **369**(9560):499-511.

14. Mignot E, Lin L, Finn L, Lopes C, Pluff K, Sundstrom ML, Young T: **Correlates of sleep-onset REM periods during the Multiple Sleep Latency Test in community adults**. *Brain* 2006, **129**(Pt 6):1609-1623.

15. Gerashchenko D, Kohls MD, Greco M, Waleh NS, Salin-Pascual R, Kilduff TS, Lappi DA, Shiromani PJ: **Hypocretin-2-saporin lesions of the lateral hypothalamus produce narcoleptic-like sleep behavior in the rat**. *J Neurosci* 2001, **21**(18):7273-7283.

16. Gerashchenko D, Murillo-Rodriguez E, Lin L, Xu M, Hallett L, Nishino S, Mignot E, Shiromani PJ: **Relationship between CSF hypocretin levels and hypocretin neuronal loss**. *Exp Neurol* 2003, **184**(2):1010-1016.

17. Echan LA, Tang HY, Ali-Khan N, Lee K, Speicher DW: **Depletion of multiple high-abundance proteins improves protein profiling capacities of human serum and plasma**. *Proteomics* 2005, **5**(13):3292-3303.

18. Greenough C, Jenkins RE, Kitteringham NR, Pirmohamed M, Park BK, Pennington SR: **A method for the rapid depletion of albumin and immunoglobulin from human plasma**. *Proteomics* 2004, **4**(10):3107-3111.

19. Gronwall C, Sjoberg A, Ramstrom M, Hoiden-Guthenberg I, Hober S, Jonasson P, Stahl S: **Affibody-mediated transferrin depletion for proteomics applications**. *Biotechnology journal* 2007, **2**(11):1389-1398.

20. Granger J, Siddiqui J, Copeland S, Remick D: **Albumin depletion of human plasma also removes low abundance proteins including the cytokines**. *Proteomics* 2005, **5**(18):4713-4718.

21. Villanueva J, Shaffer DR, Philip J, Chaparro CA, Erdjument-Bromage H, Olshen AB, Fleisher M, Lilja H, Brogi E, Boyd J *et al*: **Differential exoprotease activities confer tumor-specific serum peptidome patterns**. *J Clin Invest* 2006, **116**(1):271-284.

22. Petricoin E, Wulfkuhle J, Espina V, Liotta LA: **Clinical proteomics: revolutionizing disease detection and patient tailoring therapy**. *J Proteome Res* 2004, **3**(2):209-217.

23. Cojocel C, Maita K, Baumann K, Hook JB: **Renal processing of low molecular weight proteins**. *Pflugers Arch* 1984, **401**(4):333-339.

24. Dennis MS, Zhang M, Meng YG, Kadkhodayan M, Kirchhofer D, Combs D, Damico LA: **Albumin binding as a general strategy for improving the pharmacokinetics of proteins**. *J Biol Chem* 2002, **277**(38):35035-35043.

25. Hortin GL, Sviridov D, Anderson NL: **High-abundance polypeptides of the human plasma proteome comprising the top 4 logs of polypeptide abundance**. *Clinical chemistry* 2008, **54**(10):1608-1616.

26. Lowenthal MS, Mehta AI, Frogale K, Bandle RW, Araujo RP, Hood BL, Veenstra TD, Conrads TP, Goldsmith P, Fishman D *et al*: **Analysis of albumin-associated peptides and proteins from ovarian cancer patients**. *Clinical chemistry* 2005, **51**(10):1933-1945.

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

28. Mehta AI, Ross S, Lowenthal MS, Fusaro V, Fishman DA, Petricoin EF, 3rd, Liotta LA: **Biomarker amplification by serum carrier protein binding**. *Dis Markers* 2003, **19**(1):1-10.

29. Zhou M, Lucas DA, Chan KC, Issaq HJ, Petricoin EF, 3rd, Liotta LA, Veenstra TD, Conrads TP: **An investigation into the human serum "interactome"**. *Electrophoresis* 2004, **25**(9):1289-1298.

30. Tirumalai RS, Chan KC, Prieto DA, Issaq HJ, Conrads TP, Veenstra TD: **Characterization of the low molecular weight human serum proteome**. *Mol Cell Proteomics* 2003, **2**(10):1096-1103.

31. Bergquist J, Palmblad M, Wetterhall M, Hakansson P, Markides KE: **Peptide mapping of proteins in human body fluids using electrospray ionization Fourier transform ion cyclotron resonance mass spectrometry**. *Mass spectrometry reviews* 2002, **21**(1):2-15.

32. Lopez MF, Mikulskis A, Kuzdzal S, Bennett DA, Kelly J, Golenko E, DiCesare J, Denoyer E, Patton WF, Ediger R *et al*: **High-resolution serum proteomic profiling of Alzheimer disease samples reveals disease-specific, carrier-protein-bound mass signatures**. *Clinical chemistry* 2005, **51**(10):1946-1954.

33. Lopez MF, Mikulskis A, Kuzdzal S, Golenko E, Petricoin EF, 3rd, Liotta LA, Patton WF, Whiteley GR, Rosenblatt K, Gurnani P *et al*: **A novel, high-throughput workflow for discovery and identification of serum carrier protein-bound peptide biomarker candidates in ovarian cancer samples**. *Clinical chemistry* 2007, **53**(6):1067-1074.

34. Fisher WG, Rosenblatt KP, Fishman DA, Whiteley GR, Mikulskis A, Kuzdzal SA, Lopez MF, Tan NC, German DC, Garner HR: **A robust biomarker discovery pipeline for**

high-performance mass spectrometry data. *Journal of bioinformatics and computational biology* 2007, **5**(5):1023-1045.

35. Zimmerman LJ, Wernke GR, Caprioli RM, Liebler DC: **Identification of protein fragments as pattern features in MALDI-MS analyses of serum**. *Journal of proteome research* 2005, **4**(5):1672-1680.

36. Sturgeon CM, Hoffman BR, Chan DW, Ch'ng SL, Hammond E, Hayes DF, Liotta LA, Petricoin EF, Schmitt M, Semmes OJ *et al*: **National Academy of Clinical Biochemistry Laboratory Medicine Practice Guidelines for use of tumor markers in clinical practice: quality requirements**. *Clinical chemistry* 2008, **54**(8):e1-e10.

37. Tejler L, Grubb AO: **A complex-forming glycoprotein heterogeneous in charge and present in human plasma, urine, and cerebrospinal fluid**. *Biochimica et biophysica acta* 1976, **439**(1):82-94.

38. Matsuzaki H, Kobayashi H, Yagyu T, Wakahara K, Kondo T, Kurita N, Sekino H, Inagaki K, Suzuki M, Kanayama N *et al*: **Plasma bikunin as a favorable prognostic factor in ovarian cancer**. *J Clin Oncol* 2005, **23**(7):1463-1472.

39. Fries E, Blom AM: **Bikunin--not just a plasma proteinase inhibitor**. *The international journal of biochemistry & cell biology* 2000, **32**(2):125-137.

40. Kobayashi H: **Endogenous anti-inflammatory substances, inter-alpha-inhibitor and bikunin**. *Biological chemistry* 2006, **387**(12):1545-1549.

41. Franck C, Pedersen JZ: **Trypsin-inhibitory activities of acid-stable fragments of the inter-alpha-trypsin inhibitor in inflammatory and uraemic conditions**. *Scandinavian journal of clinical and laboratory investigation* 1983, **43**(2):151-155.

42. Sugiki M, Maruyama M, Yoshida E, Sumi H, Mihara H: **Acid-stable protease inhibitor in chronic phase of carrageenin-induced inflammation in rats**. *Inflammation* 1991, **15**(4):281-289.

43. Pugia MJ, Valdes R, Jr., Jortani SA: **Bikunin (urinary trypsin inhibitor): structure, biological relevance, and measurement**. *Advances in clinical chemistry* 2007, **44**:223-245.

44. Takano M, Mori Y, Shiraki H, Horie M, Okamoto H, Narahara M, Miyake M, Shikimi T: **Detection of bikunin mRNA in limited portions of rat brain**. *Life sciences* 1999, **65**(8):757-762.

45. Shikimi T, Hattori K, Takaori S: **Existence of a human urinary trypsin inhibitor (urinastatin)-like substance in the rat brain**. *Japanese journal of pharmacology* 1992, **60**(2):97-103.

46. Yoshida E, Sumi H, Maruyama M, Tsushima H, Matsuoka Y, Sugiki M, Mihara H: **Distribution of acid stable trypsin inhibitor immunoreactivity in normal and malignant human tissues**. *Cancer* 1989, **64**(4):860-869.

47. Yoshida E, Yoshimura M, Ito Y, Mihara H: **Demonstration of an active component of inter-alpha-trypsin inhibitor in the brains of Alzheimer type dementia**. *Biochemical and biophysical research communications* 1991, **174**(2):1015-1021.

CHAPTER FIVE
High-Throughput Evaluation of Relative Cell Permeability between Peptoids and Peptides

5.1 INTRODUCTION

In Chapter 2, a high-throughput methodology was developed and implemented for biomarker (drug target) discovery. Its application in the study of multiple sclerosis (Chapter 3) and narcolepsy (Chapter 4) demonstrated its robustness in discovering disease-specific markers that are potential therapeutic targets. This chapter describes a high-throughput study that compares the cell permeability of peptoids to peptides and show that peptoids represent a class of molecules that is suitable for drug development to target disease biomarkers.

5.1.1 General introduction

Peptides are excellent ligands for proteins and, in particular, are often capable of targeting regions of proteins not easily recognized by traditional small molecules, such as protein interaction surfaces. However, the practical utility of peptides as drugs or tools for chemical biology is limited by their sensitivity to proteases and their lack of cell permeability. Therefore, there is considerable interest in the development of compound classes with the protein-binding properties of peptides, but more favorable pharmacokinetic properties. Towards this goal, we and others have demonstrated that peptoid [1] (oligo-N-substituted glycines) libraries [2, 3] are excellent sources of protein-binding ligands [4-7]. As expected, peptoids are not sensitive to proteases or peptidases [8].

We have reported previously a cell-based assay in which entry of a peptoid- or peptide-steroid conjugate triggers the expression of a reporter gene in a dose-dependent fashion [9], allowing one to compare the relative cell permeability of various peptide- or peptoid-steroid conjugates [9]. We have found that the movement of these compounds into cells requires passive diffusion and does not appear to involve any form of active transport [10]. Using this assay, we

demonstrated that many peptoid tetramers from a combinatorial library are quite cell permeable, in that the peptoid-steroid conjugate induces reporter gene expression in the cell-based assay almost as well as the steroid alone. In other words, many tetrameric peptoids do not diminish the cell permeability of the attached steroid. Many octameric peptoids were also found to be cell permeable by this criterion, though less so than the tetramers, as expected.

We have also employed this assay to carry out careful comparisons of the relative cell permeability of a small number of isomeric peptides and peptoids. These experiments, which involved titration of the indicator cells with different concentrations of the peptide- and peptoid-steroid conjugates, confirmed that peptoids are anywhere from 3- to 30-fold more permeable than the analogous peptide, depending on the size of the molecule [11].

We were curious to determine if this conclusion was valid in general for peptoids and peptides with a wide diversity of side chains. In this report, we employ the cell-based permeability assay in a high-throughput mode to address this issue. The large number of compounds employed in this study precluded carrying out titrations and demanded a single-point analysis for each compound. Here, we address the technical issue inherent in carrying out such a screen and also consider various physical models to rationalize the results. In general, the data support the idea that peptoids are generally more cell permeable than peptides and that this difference can be attributed largely to the absence of the highly polar N–H main chain bond in peptoids.

5.2 MATERIALS AND METHODS

5.2.1 Reagents and instrumentation

All chemicals and reagents in organic synthesis were purchased from Sigma-Aldrich. For library synthesis, Polystyrene A-RAM macrobeads (500-560 μm, 0.55 mmol/g) were from Rapp Polymere. Rink amide AM resins (200-400 mesh, capacity: 0.71 mmol/g) were from NOVAbiochem. For solid phase synthesis, N-(9-Fluorenylmethoxy-carbonyl)-acetylethyleneglycol-ethyl-amine (Fmoc-AEEA-OH), O-(7-azabenzo-triazol-1-yl)-1, 1, 3, 3-tetramethyluronium hexafluorophosphate (HATU), and 1-hydroxy-7-azabenzotriazole (HoAt) were from Applied Biosystems. 2-(1H-benzotriazole-1-yl)-1, 1, 3, 3-tetramethyluronium hexafluorophosphate (HBTU) and N- hydroxybenzotriazole (HoBt) were from SynPep. Diisopropylethylamine (DIPEA) and 2, 6-lutidine were from Sigma-Aldrich. All Fmoc amino acid monomers were from SynPep and Advanced NOVAbiochem. O-tert-butyl ethanolamine was from CSPS Pharmaceuticals. Glycine tert-butyl ester acetate was from NOVAbiochem. Diaminobutane, isobutylamine, (R)-methylbenzylamine, bromoacetic acid, and diisopropylcarbodiimide (DIC) were from Sigma-Aldrich. Cell culture media and transfection reagents were purchased from Invitrogen. Preparative HPLC was performed on a Waters Binary HPLC system with a C18 reverse-phase column with the gradient elution of water/acetonitrile with 0.1 % trifluoroacetic acid (TFA). Mass spectrometry (MALDI TOF) was performed on a Voyager-DE PRO biospectrometry workstation (Applied Biosystems) with CHCA as matrix.

5.2.2 Syntheses of OxDex, SDex, SDex-Peptoid and SDex-Peptide analogs

OxDex-COOH (Dex-17β-carboxylic acid) and SDex-COOH (Dex-21-thiopropionic acid) were synthesized based on previously published procedures [9, 11]. SDex-conjugated peptoid and peptide analogs were synthesized as described previously [11]. The analogs were capped with SDex-COOH steroid.

5.2.3 Syntheses of OxDex-Peptoid and OxDex-Peptide libraries

Both peptoid and peptide libraries were constructed on Polystyrene A-RAM macrobeads (500-560 μm, 0.55 mmol/g) from Rapp Polymere. Peptoid library synthesis was performed as described previously with slight modifications [2]. The synthesis of peptoids under microwave conditions was performed in a 1000 W Whirlpool microwave oven (model MT1130SG) with 10% power. Peptide library synthesis was performed using standard solid phase synthesis methods. The synthesis of peptides was performed in a New Brunswick Scientific Innova 4000 incubator shaker. Standard glass peptide synthesis vessels (Chemglass) were used for the synthesis in the incubator shaker and in the microwave oven. Upon completion of the library syntheses, Fmoc-AEEA-OH and OxDex-COOH were coupled to the beads using standard solid phase synthesis methods. The libraries were sorted into 96-well plates in a one bead per well fashion and cleaved with trifluoroacetic acid:water (95:5 vol/vol) at room temperature with slight agitation for 2 hours. The TFA was evaporated under hoods and the compounds were resuspended in 20 μl 50% acetonitrile in water. One-fourth (5 μl) of the compounds per bead was aliquoted into a separate 96-well plate for sequencing purposes. The remaining three-fourths (~60nmol) of compounds were dried and resuspended in 11.5 μl 10% DMSO in water for cell culture experiments. About 10 nmol of each compound was used so that each well contained about 100 μM of each OxDex-capped molecule in 100 μl of cell culture media.

5.2.4 Plasmids, cell culture, transfection, *in vitro* competition GR binding assays and high-throughput cell permeability luciferase assay

Procedures were performed as described previously [9-11]. HeLa cells were grown in 96-well plates for the high-throughput study.

5.2.5 Permeability ratio determination

The permeability ratio (PR) was obtained by dividing the luminescence reading from the firefly luciferase activity to the luminescence reading from the *Renilla* luciferase activity. The PR was then normalized to the negative (no compound) and positive (dexamethasone) control ratios from each plate. PR = (Firefly luminescence / *Renilla* luminescence) / (No compound luminescence / Dexamethasone luminescence).

5.2.6 Statistical analysis

The mean luminescence reading for the internal control *Renilla* luciferase was obtained from all 8 plates. Compounds with a *Renilla* luminescence reading that lies greater than ± 2 standard deviations from the mean were omitted from the study. The mean normalized PR for each class of compounds was calculated from the remaining compounds. The 2-tail t-test on mean PR between peptoids and peptides was performed using the statistical software SAS.

5.2.7 Peptoid and peptide sequencing

Compounds that have a normalized PR that is at least 2 standard deviations greater than the mean ratio were analyzed using the Applied Biosystems 4700 Proteomics Analyzer with TOF/TOF optics to obtain the molecular ion mass. The possible monomeric composition of the molecule was predicted from an in-house program written in Perl using the molecular ion mass, the known generic structure of the libraries, and the residue mass of the monomers. Tandem mass spectrometry sequencing on the same instrument was performed to confirm the sequences.

5.2.8 Physicochemical property computations

All physicochemical property calculations were obtained using Molinspiration Cheminformatics (http://www.molinspiration.com/cgi-bin/properties). LogP prediction is based on group contributions and takes into account the intramolecular hydrogen bonding contribution to logP and charge interactions. TPSA calculation is based on the summation of tabulated surface contributions of polar fragments (atoms regarding also their environment) and provides results of practically the same quality as the classical 3D PSA, but is two to three orders of magnitude faster [12].

5.3 RESULTS AND DISCUSSION

5.3.1 Library design and synthesis

Our previous studies with steroid-conjugated peptoid and peptide analogs demonstrated that as the length of the molecule increases from a dimer to an octamer, the cell permeability decreases [9-11]. Based on this observation, we decided to synthesize libraries of tetramers for our current high-throughput comparison study as a compromise between maintaining the diversity of the libraries and retaining decent permeability within the molecules. The peptoid libraries were synthesized using the 'sub-monomer' synthesis [3] in a conventional split-and-pool approach. The amines employed in this chemistry contained cationic, anionic, hydrophobic, and neutral moieties (Panel A, Fig. 5.1). The peptide libraries displayed comparable side chains (Panel B, Fig. 5.1) and were constructed by standard solid-phase Fmoc chemistry. High-capacity polystyrene macrobeads were used to generate a sufficient amount of compound per bead (~80 nmol) for the cell permeability assay upon cleavage from the beads, in a one bead per well manner.

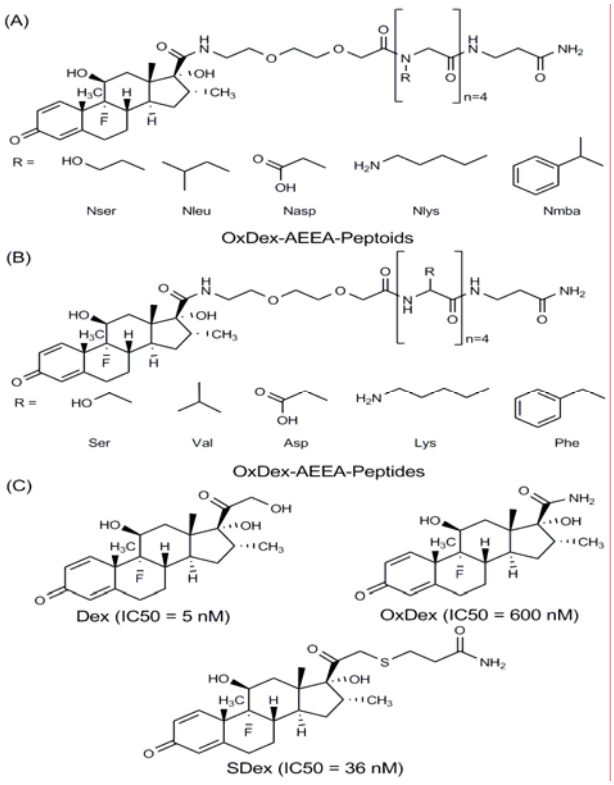


Figure 5.1 **Chemical entities used in high-throughput study**. Generic chemical structure of the OxDex-conjugated (A) peptoid and (B) peptide libraries and the side chain moieties incorporated. (C) Chemical structure of dexamethasone and its amine derivatives, OxDex and SDex.

The generic structure of the libraries was OxDex-AEEA-X4-β-Ala, where β-alanine is the invariant C-terminal residue, and Ox-Dex is the steroid capping molecule at the N-terminus conjugated to the four variant monomers (X) via an acetyl-ethyleneglycolethyl-amine (AEEA) linker (Panels A and B, Fig. 5.1). OxDex is product of oxidative cleavage of the dexamethasone side chain [13]. Panel C in Figure 5.1 shows the amide form of OxDex used to determine its IC_{50} value, a measure of relative binding affinity to the glucocorticoid receptor (GR). Quality control analyses were performed on both libraries by subjecting randomly selected beads from each library to high-performance liquid chromatography (HPLC), followed by mass spectrometry

(MS) analysis. Nine of the ten beads showed a clear major product peak in the HPLC traces with 90% purity and a single dominant peak was observed in the mass spectra in the expected mass range. Note that OxDex was linked to the peptide or peptoid during solid-phase synthesis, and uncoupled steroid was removed by thorough washing. Thus, the solutions are not contaminated by free steroid, which would, of course, skew the results of the permeability assays. Furthermore, we have shown previously that the steroid-peptoid or -peptide linkage is stable in cell culture medium [10, 11].

5.3.2 High-throughput cell permeability assay

The high-throughput cell permeability assay employed in this study has been described elsewhere [9, 10]. A schematic of the system is depicted in Figure 5.2. Briefly, the OxDex conjugates were exposed to HeLa cells transfected with three plasmids. One encodes for a fusion protein containing the Gal4 DNA-binding domain, the GR ligand-binding domain and the VP16 transactivation domain (Gal4DBD-GRLBD-VP16). The apo form of this protein is sequestered in the cytoplasm in its inactive form through a tight interaction with heat shock protein 90 (Hsp90) in the absence of ligand This interaction is disrupted by an influx of the steroid, which binds to the GR LBD, allowing the fusion protein to translocate into the nucleus. It then drives the expression of firefly luciferase expression by activating the Gal4-responsive firefly luciferase reporter gene carried by the second plasmid. We have shown that the affinity of different peptide- and peptoid-steroid complexes for the GR LBD differs only modestly, thanks to the presence of the β-alanine linker between the variable sequence and the steroid [9, 10]. Therefore, expression of the firefly luciferase is dependent on the permeability of the steroid-conjugated molecule. The third plasmid carries a constitutively expressed Renilla reniformis luciferase gene that serves as a transfection control. The ratio of firefly luciferase activity (compound-dependent) to Renilla luciferase activity (compound-independent internal control) is a reflection of the concentration of

steroid conjugates that have successfully permeated the cell membrane into the cell. This ratio compensates for the well-to-well variability that could potentially arise during readouts. Four 96-well plates from each library were used in this comparison study. To address the possibility of plate-to-plate variability, the permeability ratio of each compound in each plate was normalized to the plate's ratio of the negative (no compound) and positive (dexamethasone) controls.

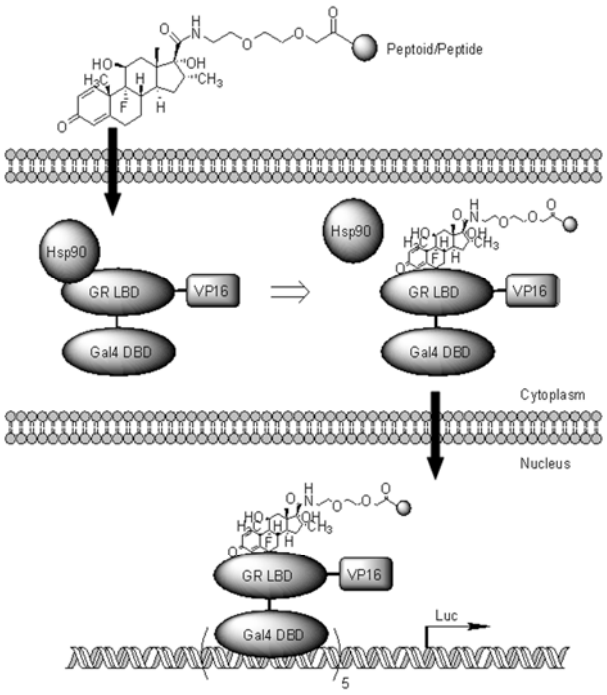


Figure 5.2 **Schematic illustration of the cell permeability assay used in the high-throughput study of steroid conjugates**. Refer to text for details.

5.3.3 Technical issues

A total of eight 96-well plates (four from each library) were assayed simultaneously using the same batch of HeLa cells to minimize variability due to cell preparation. For practical reasons, a Tecan SpectraFluor Plus plate reader was chosen for the luminescence activity readout due to the high-throughput nature of the study. Titration of steroid-conjugated peptoid and peptide analogs for comparison between the Tecan and Sirius luminometer (Berthold Detection Systems) revealed that the reduced sensitivity of the Tecan resulted in lower luminescence readout values than the luminometer. Fortunately, the Tecan plate reader preserved the relative difference between the molecules and produced comparable EC_{50} values (luciferase induction) for the same molecules, albeit with a reading two orders of magnitude lower than that of the single-tube luminometer.

In the actual study, the final concentration of OxDex-conjugated peptoid or peptide exposed to HeLa cells in each well was about 100 μM since the construct OxDex-AEEA-$CONH_2$ was shown to have an EC_{50} value of <50 μM [9]. Therefore, at this concentration, we expect the OxDex conjugates to show at least half-maximum induction of luciferase expression, allowing us to clearly distinguish between permeable and impermeable compounds. Because of the large number of molecules analyzed in this study, it was not feasible to examine multiple compound concentrations to obtain a titration curve. Therefore, it was imperative to determine beforehand that the readouts will be in the linear, and not the saturated, part of such a titration curve. This appeared to be the case for two particular peptoids that had been identified previously as having relatively high cell permeability [9]. However, for two peptoids that had been previously identified as being poorly cell permeable, only a small amount of reporter gene induction was observed at an extracellular concentration of 100 μM and it is not clear if this value is in the linear range. Based on this observation, we assume that all of the single-point readings for peptoids and peptides that score as relatively permeable reflect the linear part of the titration curve for each compound, but that this assumption may not be valid for some peptoids that are

unusually cell impermeable. Of course, another complicating factor is that in the analysis of the library, we cannot be certain that the concentration of each peptoid- or peptide-steroid conjugate is the same due to possible differences in synthesis efficiency, though we believe that these differences are not large, based on the aforementioned analysis of several compounds chosen randomly from the library.

5.3.4 Comparison of peptoids and peptides

The current study compared the relative cell permeability of 350 steroid-conjugated peptoids and 350 steroid-conjugated peptides. The permeability ratio for each compound per well was calculated as (Firefly luciferase luminescence/*Renilla* luciferase luminescence)/(Blank negative control luminescence/ Dexamethasone positive control luminescence). The average permeability ratio from each class of molecules was subject to a 2-tail t-test using SAS. The statistical analysis showed that the two groups of compounds passed the equality of variance requirement and had a statistically significant difference with a p-value of <0.01 when the average permeability ratio of peptoids (0.0118) was compared to the average permeability ratio of peptides (0.0059) (Fig. 5.3).

The relative permeability investigated here obviously reflects that of the steroid conjugates and may not reflect the true permeability of the parent compounds. However, any assay involving labeling of the molecule of interest suffers from this limitation, including confocal fluorescence microscopy. Thus, we restrict our comments in the discussion below to relative statements comparing one set of steroid-substituted molecules to another.

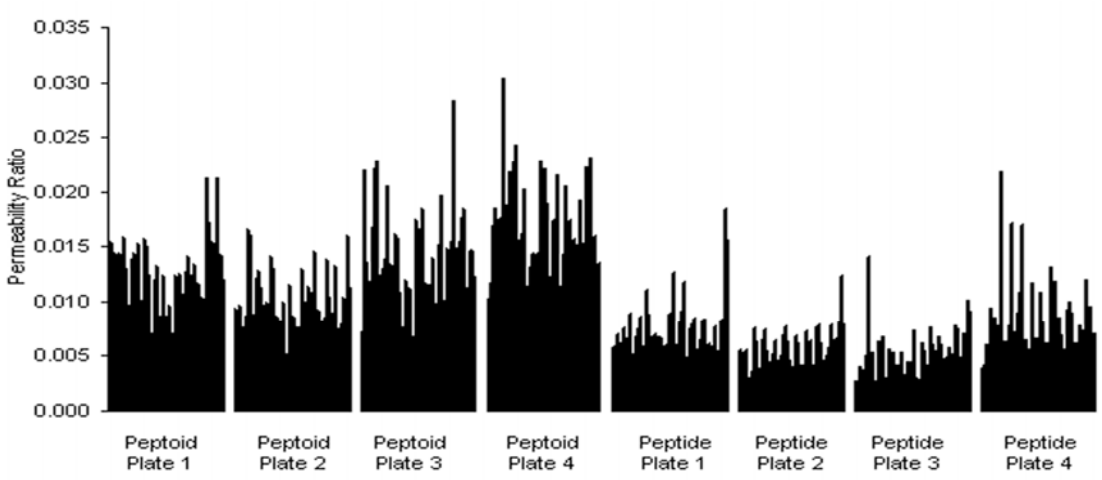


Figure 5.3 **High-throughput comparison of relative cell permeability between peptoids and peptides**. Permeability ratio = (Firefly luciferase luminescence / Renilla luciferase luminescence) / (Blank negative control luminescence / Dexamethasone positive control luminescence).

5.3.5 Comparison of physicochemical properties with permeability

The prediction of absorption, distribution, metabolism, and excretion (ADME) properties of organic molecules continue to play a critical role in the drug discovery process and it is now possible to make reasonable estimates of the permeability of organic compounds based on their molecular structures. In an attempt to delineate the factors that are responsible for the observed difference in permeability between peptoids and peptides, we investigated a number of physicochemical properties that have been shown to affect permeability [12, 14, 15]. We first considered highly cell permeable peptoids and peptides that have a permeability ratio at least two standard deviations greater than the average permeability ratio of its class. The composition of the peptoids and peptides that scored as such was identified by mass spectrometry.

The physicochemical parameters examined include lipophilicity, polar surface area (PSA), hydrogen-bonding capacity (hydrogen bond acceptors and donors), molecular size, molecular volume, and molecular rigidity, which were calculated using Molinspiration Cheminformatics as reported elsewhere [12].

5.3.5.1 Lipophilicity

A traditional molecular transport descriptor that is used to model permeability is the n-octanol–water partition coefficient ($logP_{oct}$ or simply logP). LogP is still widely used as a measure of hydrophobicity or lipophilicity that affects membrane penetration and permeability [14-17], with a higher logP value indicating greater lipophilicity. The computational logP (ClogP) values in this study take into consideration the intramolecular hydrogen-bonding contribution and charge interactions. In our study, peptoids appear more cell permeable and tend to have a lower ClogP value than peptides, suggesting peptoids are slightly less lipophilic than peptides. The average ClogP value of peptoids in the high-throughput study is lower than that of peptides (-2.67 vs -1.90) (Table 5.1). This trend of lower ClogP value with an observed higher permeability is in line with the study of steroid transport across Caco-2 monolayer cells by Faassen and co-workers [18] and other work [19].

Compound	PR	logP	MW	Nrotb	Molecular volume ($Å^3$)
Peptoids	0.0221	-2.67	1061.65	30	976.69
Peptides	0.0134	-1.90	1074.00	29	985.49

Table 5.1. **Mean value of selected molecular descriptors in high-throughput study**. PR is the mean permeability ratio obtained from the highly permeable OxDex-conjugated compounds ($\geq 2\sigma$ above average permeability ratio, n=12 for peptoids and n=12 for peptides). MW represents the molecular weight. Nrotb is the number of rotatable bonds.

This phenomenon could be attributed to the peptoid residues, which as imino acids are more hydrophilic than the corresponding amino acid counterparts based on the Liu–Deber hydropathy scale [20]. It is known that excessive lipophilicity is a common cause of poor solubility [21] and thus leads to poor cell permeability. In converting the residue to an N-alkylglycine moiety, the N-atom becomes tertiary and more basic, likely increasing the water solubility of the peptoids. Although a good predictor of permeation across biological membranes, lipophilicity is not the sole determinant of cell permeability and other factors have been shown to play a role in affecting the overall permeability of a molecule.

5.3.5.2 Polar surface area (PSA)

In recent years, the PSA of a molecule has emerged as a key predictor of permeability. PSA is defined as the sum of the van der Waals (or solvent-accessible) surface areas of oxygen and nitrogen atoms, including attached hydrogens [22]. The topological polar surface area (TPSA) was calculated to investigate this parameter in our study. It is computationally up to three times faster and comparable to the classical 3D PSA. In our high-throughput study, the average TPSA value of the highly permeable peptoids of 335.30 $Å^2$ is also lower than the highly permeable peptides' average TPSA value of 358.80 $Å^2$ (Table 5.2). This is a relatively small difference and may play only a modest role in the observed differences in the permeability of peptides and peptoids. To the extent that it does contribute, this means that overall, peptoids have less polar groups exposed to solvents than peptides, which in turn suggests that the resulting lower TPSA value might contribute to the higher cell permeability of peptoids over peptides. This trend is also present within the OxDex-conjugated peptoids with differing permeabilities. Peptoids with a lower TPSA tend to have a higher permeability ratio (Fig. 5.4).

Compound	PR	TPSA (Å^2)	H-bond acceptors	H-bond donors	Total H-bonds
Peptoids	0.0221	335.30	23	9	32
Peptides	0.0134	358.80	22	14	36

Table 5.2. **Hydrogen bonding capacity parameters in high-throughput study**. PR is the mean permeability ratio obtained from the highly permeable OxDex-conjugated compounds (≥ 2σ above mean permeability ratio, n=12 for peptoids and n=12 for peptides). Total H-bonds is the sum of H-bond acceptors and donors.

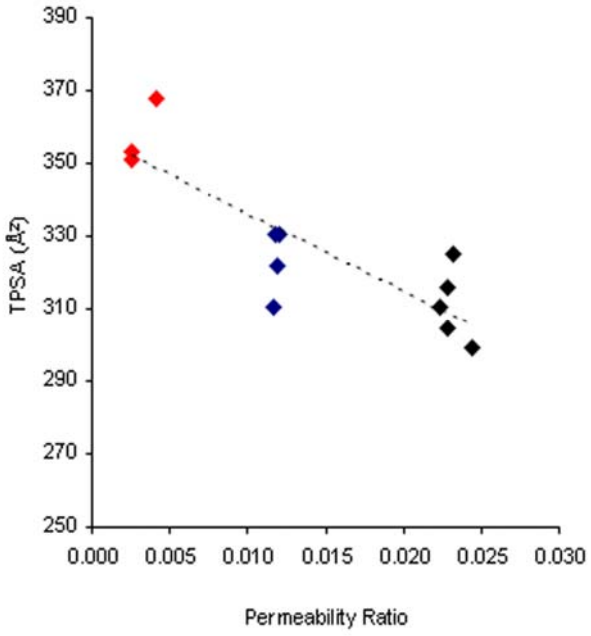


Figure 5.4. **Trend in TPSA in OxDex-conjugated peptoids**. Peptoids with permeability ratio at least 2σ below (red), above (black), or at (blue) mean permeability ratio value.

Literature suggests that permeability is optimal when PSA is <120 $Å^2$ based on a study of commercially available drugs [23]. Although all the molecules in our study have a PSA > 120 $Å^2$, the trend of lower PSA conferring higher permeability is observed, though the effect is subtle. PSA has been used to predict passage though the blood–brain barrier (BBB), flux across Caco-2 monolayers, and human intestinal absorption [22-24], and succeeded in providing good correlation with experimental transport data [12]. In our study, the lower TPSA in peptoids can be attributed to the conversion of the backbone amide from a secondary to a tertiary nitrogen which eliminates the very polar amide bond found in peptides. Consequently, the tendency for a hydration shell to form around the peptoid is reduced. It is postulated that polar groups resist desolvation when they move from an aqueous extracellular environment to the more lipophilic interior of membranes. The PSA thus may reflect at least part of the desolvation energy for breaking the solute:water interaction necessary in membrane transport. Specifically, the higher PSA in peptides may indicate the greater desolvation energy required to overcome the strong amide:water interactions in peptides.

5.3.5.3 Hydrogen bonding capacity

The polar functionalities of the PSA parameter of a compound can be related to its hydrogen-accepting and hydrogen-donating ability, with hydrogen-bonding being one of two main components of lipophilicity. The peptoids and peptides have an equal number of hydrogen acceptors but the peptoids have a lower number of hydrogen donors (average of 9 hydrogen donors in peptoids against 14 in peptides) (Table 5.2), resulting in a reduced total hydrogen-bonding capacity (sum of hydrogen acceptors and donors). This correlated with increased permeability. A modest trend of this sort is observed within the OxDex-conjugated peptoids (Fig. 5.5).

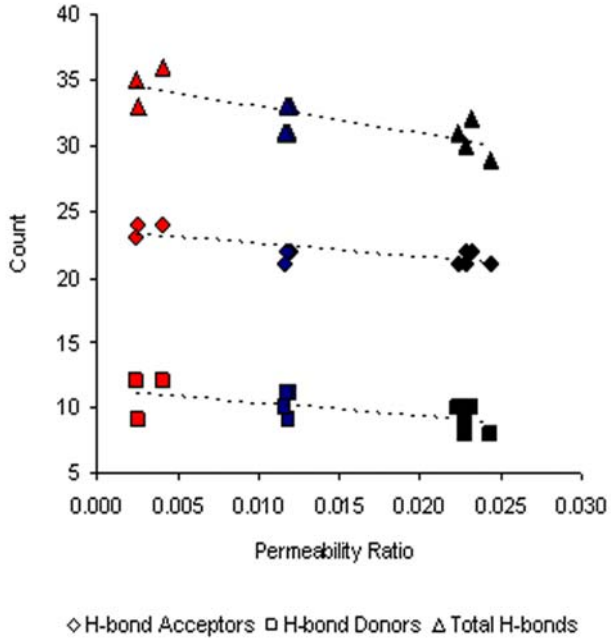


Figure 5.5. **Trend in hydrogen bonding capacity in OxDex-conjugated peptoids**. Peptoids with permeability ratio at least 2σ below (red), above (black), or at (blue) mean permeability ratio value. Total H-bonds is the sum of H-bond acceptors and donors.

Since the structural difference between peptoids and peptides mainly affects the hydrogen-bonding potential of peptoids, we further investigated whether it could be the main physicochemical parameter responsible for the higher permeability seen in peptoids. Highly permeable molecules from both classes with comparable ClogP values were compared for differences in hydrogen-bonding capacity. We found that the TPSA and hydrogen bonding capacity were decreased in peptoids, specifically in a reduction of hydrogen bond donor potential (Table 5.3). This reduction that accompanies higher permeability seen in peptoids supports the notion that lipophilicity can affect cell permeability only to a certain extent whereas the hydrogen-bonding potential might assume a more determinant role, as suggested previously [25].

Compound	PR	logP	TPSA (Å^2)	H-bond acceptors	H-bond donors	Total H-bonds
Peptoids	0.0229	-1.70	319.03	22	9	31
Peptides	0.0134	-1.38	353.71	22	13	35

Table 5.3. **Comparison of logP and hydrogen bonding capacity parameters in high-throughput study**. PR is the mean permeability ratio obtained from the highly permeable OxDex-conjugated compounds ($\geq 2\sigma$ above mean permeability ratio) with comparable logP values (n=7 for peptoids and n=6 for peptides). Total H-bonds is the sum of H-bond acceptors and donors.

Hydrogen-bonding capacity bears such significance that it constitutes two of Lipinski's Rule of Five [26] of drug design. It has been found that the hydrogen-bonding capacity of a drug solute correlates reasonably well with passive diffusion [27, 28]. An increased N–H bond count for both acceptors and donors tends to worsen permeability [29, 30]. To further substantiate this argument, compounds with high hydrogen forming potential, such as peptides with their amide groups as small as di- and tripeptides, have minimal distribution through the BBB, while compounds possessing a tertiary nitrogen show a high degree of brain permeation [31]. Indeed, tertiary nitrogen is a feature of many central nervous system drugs [31]. Hydrogen-bonding potential might thus constitute the limiting step in cell permeation.

Molecules with very polar amide bonds like peptides have greater hydrogen-bonding interactions with the surrounding water. As a result, the desolvation energy required to break these interactions in order for the peptides to transfer from the hydrophilic, aqueous environment to the hydrophobic, non-hydrogen bonding membrane interior is substantially increased. Moreover, once the solute:water interaction is overcome, cell permeability can be further hindered by the binding of the molecules to the lipid-rich layer of cell membranes through the donation of hydrogen-bonds as it approaches the polar surface of the membrane and desolvates as

it moves into the lipid portion [32]. In fact, the hydrophilic part of lipids contains hydrogen-bonding acceptor groups which may hinder the transbilayer insertion of the high hydrogen-donating molecules and prevent their transport across a cellular membrane via tight binding. Hence, the greater hydrogen-donating potential of peptides over peptoids might be the cause of lower permeability seen in peptides. It is therefore not surprising that modifications that result in the reduction of hydrogen-bond-donating capacity, such as conformational facilitation of intramolecular hydrogen bonds [32-38] and the absence of hydrogen donors altogether from the tertiary amines in the case of peptoids, will facilitate membrane permeation.

5.3.5.4 Molecular size, volume, and rigidity

Molecular size is the second basic component of solubility and permeability. Molecular weight is a surrogate for other properties, including molecular volume and rigidity. The simplest measure of molecule rigidity is by determining the number of rotatable bonds present. In our high-throughput study, molecular size and volume are comparable between the OxDex-conjugated peptoids and peptides. Similarly, there is no difference in the number of rotatable bonds between the two classes of molecules under investigation (Table 5.1). Even though these molecular properties have been implicated in cell permeability [39], they appear to play a minimal role in the permeability difference observed between peptoids and peptides in this study.

5.3.5.5 Side chain composition

To determine if specific side chain characteristics are preferred over others in the more permeable molecules, the prevalence of each side chain used in the library was probed. The molecular formulas of the highly cell permeable molecules from both classes were predicted from their molecular ion mass obtained via mass spectrometry, the generic molecular structure, and the mass of the five residues used in the library using an in-house program. The side chains were then confirmed via tandem mass spectrometry sequencing, and the frequency of occurrences tabulated.

It appears that the highly cell permeable peptoids and peptides consist of 1.5 times more hydrophilic residues than hydrophobic ones (Panels A and B, Fig. 5.6), where hydrophilic residues are the positively charged lysine and the negatively charged aspartic acid. Phenylalanine and valine constitute the hydrophobic residues. Of the 12 highly permeable molecules from each class, a large majority of them are charged (11 for peptoids, 10 for peptides). Out of these charged molecules, seven peptoids and nine peptides have two or more charged residues.

The observation that the more cell permeable molecules tend to consist of more charged residues than hydrophobic ones corresponds well with an increasing body of evidence supporting ion partitioning over neutral molecules, as substantiated by studies suggesting ionic species may contribute significantly to transport across Caco-2 monolayers [38, 40]. Since our cell permeability assay uses steroid conjugates, the charges on the highly permeable molecules may serve to increase solubility of the molecules in the presence of the hydrophobic steroid. Amphiphilicity (the combination of the hydrophilic and hydrophobic parts of a molecule) may in itself influence cell permeability and deserves further investigation [21].

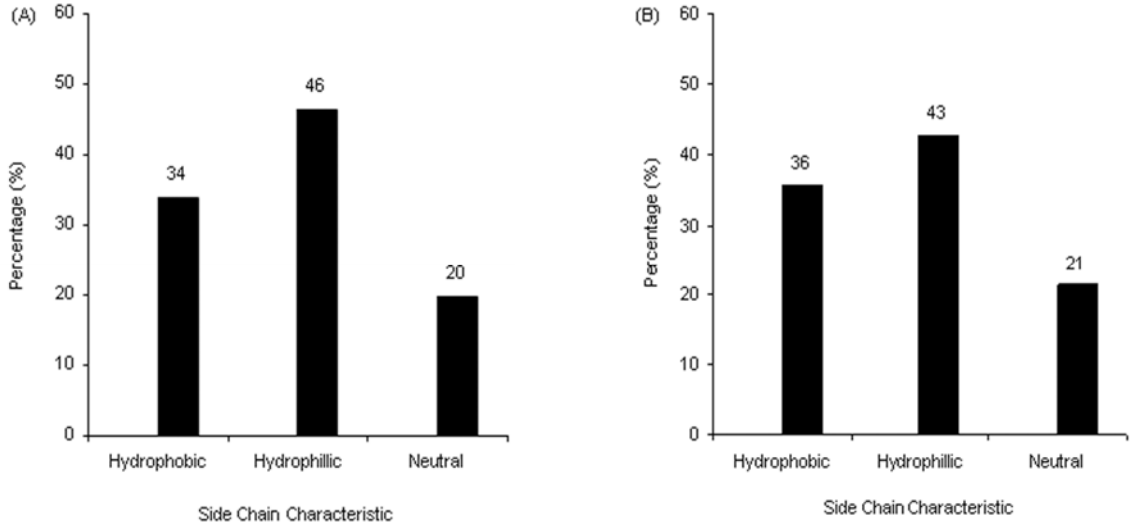


Figure 5.6 **Side chain characteristic prevalence of highly permeable (≥ 2σ above mean permeability ratio) (A) peptoids and (B) peptides**. Hydrophobic residues are Nleu and Nmba, hydrophilic residues are Ngly and Nlys, and neutral residue is Nser for peptoids. Hydrophobic residues are Val and Phe, hydrophilic residues are Asp and Lys, and neutral residue is Ser for peptides.

5.3.5.6 Physicochemical property evaluation of SDex-conjugates

A similar evaluation of the above parameters was performed on the SDex conjugates used in the study by Kwon *et al*. [11] (Fig. 5.7). SDex is a higher affinity GR ligand than is OxDex (Panel C, Fig. 5.1) and thus shifts the titration curve to lower compound concentrations. The relative permeability comparison between peptoids and peptides in this set of conjugates involved far fewer compounds but was based on careful titrations of the exact isomeric analogs. Specifically in Kwon's study, a series of peptoid and peptide conjugates containing leucine and homoserine side chains and varying in length from dimers to octamers were synthesized and their relative cell permeability compared using the same cell-based reporter assay employed in our high-throughput study [11]. It was found that peptoids were more cell permeable than peptides, with shorter

conjugates exhibiting higher permeability. The physicochemical properties discussed above for the high-throughput study of Ox-Dex conjugates were calculated for this set of SDex-conjugated analogs and their values compared between peptoids and peptides. Not surprisingly, the trends observed in the single-point readout but more extensive set of molecules in the OxDex conjugates were mirrored in the SDex-conjugated molecules (Figs. 5.8–5.10, Tables 5.4 and 5.5).

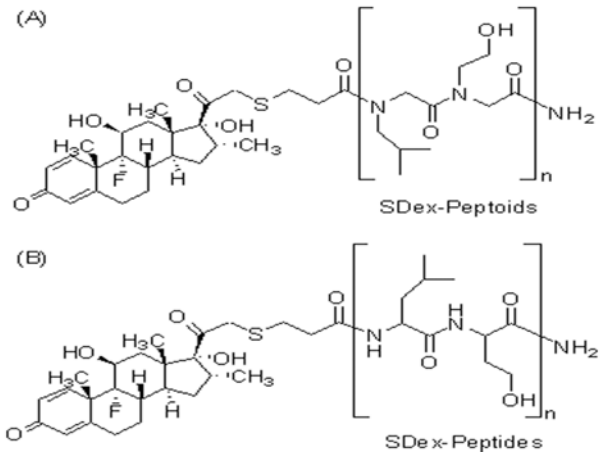


Figure 5.7 **Chemical entities used in analog study**. Generic chemical structure of the SDex-conjugated analogs of (A) peptoids and (B) peptides, where n = 1, 2, 3, or 4.

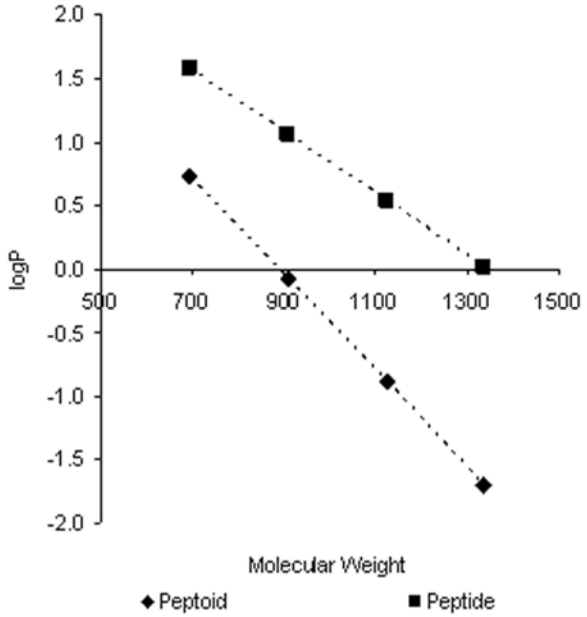


Figure 5.8. **Comparison of logP value of SDex-conjugated peptoid (POn) and peptide (PIn) analogs**. Dimers (PO2, PI2), tetramers (PO4, PI4), hexamers (PO6, PI6), and octamers (PO8, PI8) are represented by their molecular weight.

Compound	logP	Molecular weight	Number of rotatable bonds	Molecular volume (Å^3)
PO2	0.73	693.88	14	637.57
PO4	-0.08	908.14	22	843.08
PO6	-0.89	1122.41	30	1048.59
PO8	-1.70	1336.67	38	1254.10
PI2	1.58	693.88	14	636.86
PI4	1.05	908.14	22	841.66
PI6	0.53	1122.41	30	1046.46
PI8	0.01	1336.67	38	1251.26

Table 5.4 **Mean value of selected molecular descriptors in analog study**. All molecules represented here are SDex-conjugated analogs of peptoids (POn) and peptides (PIn), where molecule length n = 2, 4, 6, or 8.

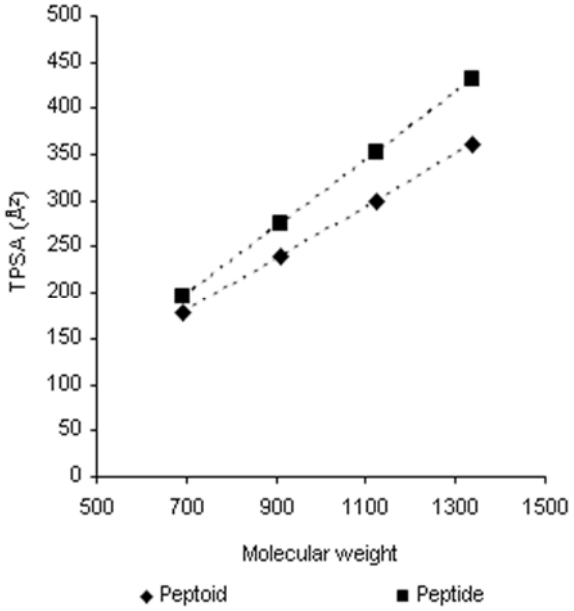


Figure 5.9 **Comparison of TPSA value between SDex-conjugated peptoid (POn) and peptide (PIn) analogs**. Dimers (PO2, PI2), tetramers (PO4, PI4), hexamers (PO6, PI6), and octamers (PO8, PI8) are represented by their molecular weight.

Compound	MW	TPSA ($Å^2$)	H-bond acceptors	H-bond donors	Total H-bonds
PO2	693.88	178.54	11	5	16
PO4	908.14	239.38	16	6	22
PO6	1122.41	300.23	21	7	28
PO8	1336.67	361.08	26	8	34
PI2	693.88	196.12	11	7	18
PI4	908.14	274.54	16	10	26
PI6	1122.41	352.96	21	13	34
PI8	1336.67	431.39	26	16	42

Table 5.5 **Hydrogen bonding capacity parameters in analog study**. All molecules represented here are SDex-conjugated analogs of peptoids (POn) and peptides (PIn), where molecule length n = 2, 4, 6, or 8. MW represents the molecular weight. Total H-bonds is the sum of H-bond acceptors and donors.

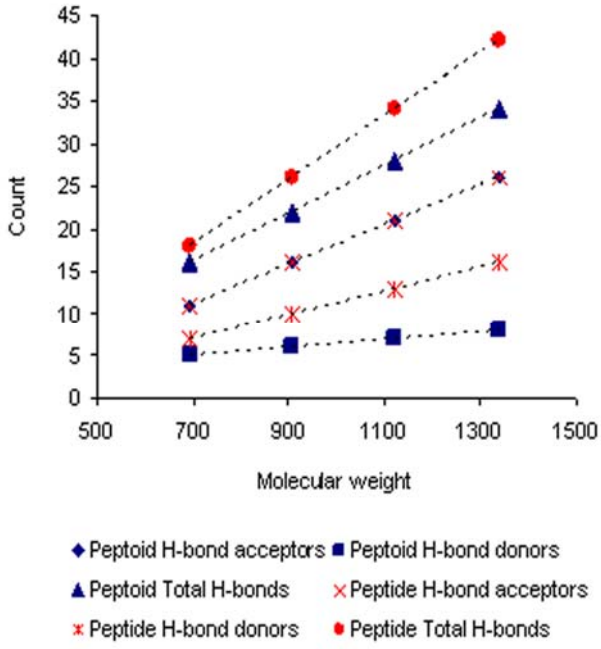


Figure 5.10 **Comparison of hydrogen bond capacity between SDex-conjugated peptoid (POn) and peptide (PIn) analogs**. Dimers (PO2, PI2), tetramers (PO4, PI4), hexamers (PO6, PI6), and octamers (PO8, PI8) are represented by their molecular weight. Total H-bonds is the sum of H-bond acceptors and donors.

5.4 CONCLUSION

Undoubtedly, the overall permeability of a molecule is determined by the delicate balance of numerous parameters that are clearly interrelated such that changing one will affect the others. To date, there exists no prediction method for this crucial attribute in drug design. In this study, we evaluated whether a cell-based permeability assay, when conducted in high-throughput mode, could contribute to helping us understand the differences in cell permeability of peptides and peptoids. As mentioned in the introduction, careful previous studies from our laboratory of a few compounds have shown that peptoids are anywhere from 3 to 30 times more permeable than comparable peptides, depending on the compound. The single-point nature of the assays

conducted here appear to flatten this difference somewhat as, on average, the peptoids were found to be about twice as permeable as the peptides. Nonetheless, the general trend held, allowing us to attempt to correlate various aspects of the molecular characteristics of some of the molecules with the observed relative cell permeability. Of special interest is the reduction in hydrogen-bond-donating potential of peptoids, which appears to be the dominant factor accounting for the increased cell permeability.

5.5 BIBLIOGRAPHY

1. Simon RJ, Kania RS, Zuckermann RN, Huebner VD, Jewell DA, Banville S, Ng S, Wang L, Rosenberg S, Marlowe CK *et al*: **Peptoids: a modular approach to drug discovery**. *Proceedings of the National Academy of Sciences of the United States of America* 1992, **89**(20):9367-9371.
2. Alluri PG, Reddy MM, Bachhawat-Sikder K, Olivos HJ, Kodadek T: **Isolation of protein ligands from large peptoid libraries**. *Journal of the American Chemical Society* 2003, **125**(46):13995-14004.
3. Figliozzi GM, Goldsmith R, Ng SC, Banville SC, Zuckermann RN: **Synthesis of N-substituted glycine peptoid libraries**. *Methods in enzymology* 1996, **267**:437-447.
4. Lim HS, Archer CT, Kodadek T: **Identification of a peptoid inhibitor of the proteasome 19S regulatory particle**. *Journal of the American Chemical Society* 2007, **129**(25):7750-7751.
5. Alluri P, Liu B, Yu P, Xiao X, Kodadek T: **Isolation and characterization of coactivator-binding peptoids from a combinatorial library**. *Molecular bioSystems* 2006, **2**(11):568-579.
6. Kodadek T, Reddy MM, Olivos HJ, Bachhawat-Sikder K, Alluri PG: **Synthetic molecules as antibody replacements**. *Accounts of chemical research* 2004, **37**(9):711-718.
7. Udugamasooriya DG, Dineen SP, Brekken RA, Kodadek T: **A peptoid "antibody surrogate" that antagonizes VEGF receptor 2 activity**. *Journal of the American Chemical Society* 2008, **130**(17):5744-5752.
8. Wang Y, Lin H, Tullman R, Jewell CF, Jr., Weetall ML, Tse FL: **Absorption and disposition of a tripeptoid and a tetrapeptide in the rat**. *Biopharmaceutics & drug disposition* 1999, **20**(2):69-75.

9. Yu P, Liu B, Kodadek T: **A high-throughput assay for assessing the cell permeability of combinatorial libraries**. *Nature biotechnology* 2005, **23**(6):746-751.

10. Yu P, Liu B, Kodadek T: **A convenient, high-throughput assay for measuring the relative cell permeability of synthetic compounds**. *Nature protocols* 2007, **2**(1):23-30.

11. Kwon YU, Kodadek T: **Quantitative evaluation of the relative cell permeability of peptoids and peptides**. *Journal of the American Chemical Society* 2007, **129**(6):1508-1509.

12. Ertl P, Rohde B, Selzer P: **Fast calculation of molecular polar surface area as a sum of fragment-based contributions and its application to the prediction of drug transport properties**. *Journal of medicinal chemistry* 2000, **43**(20):3714-3717.

13. Licitra EJ, Liu JO: **A three-hybrid system for detecting small ligand-protein receptor interactions**. *Proceedings of the National Academy of Sciences of the United States of America* 1996, **93**(23):12817-12821.

14. Young RC, Mitchell RC, Brown TH, Ganellin CR, Griffiths R, Jones M, Rana KK, Saunders D, Smith IR, Sore NE *et al*: **Development of a new physicochemical model for brain penetration and its application to the design of centrally acting H2 receptor histamine antagonists**. *Journal of medicinal chemistry* 1988, **31**(3):656-671.

15. Artursson P, Karlsson J: **Correlation between oral drug absorption in humans and apparent drug permeability coefficients in human intestinal epithelial (Caco-2) cells**. *Biochemical and biophysical research communications* 1991, **175**(3):880-885.

16. Winiwarter S, Bonham NM, Ax F, Hallberg A, Lennernas H, Karlen A: **Correlation of human jejunal permeability (in vivo) of drugs with experimentally and theoretically derived parameters. A multivariate data analysis approach**. *Journal of medicinal chemistry* 1998, **41**(25):4939-4949.

17. van de Waterbeemd H, Camenisch G, Folkers G, Chretien JR, Raevsky OA: **Estimation of blood-brain barrier crossing of drugs using molecular size and shape, and H-bonding descriptors**. *Journal of drug targeting* 1998, **6**(2):151-165.

18. Faassen F, Kelder J, Lenders J, Onderwater R, Vromans H: **Physicochemical properties and transport of steroids across Caco-2 cells**. *Pharmaceutical research* 2003, **20**(2):177-186.

19. Dhanasekaran M, Palian MM, Alves I, Yeomans L, Keyari CM, Davis P, Bilsky EJ, Egleton RD, Yamamura HI, Jacobsen NE *et al*: **Glycopeptides related to beta-endorphin adopt helical amphipathic conformations in the presence of lipid bilayers**. *Journal of the American Chemical Society* 2005, **127**(15):5435-5448.

20. Deber CM, Wang C, Liu LP, Prior AS, Agrawal S, Muskat BL, Cuticchia AJ: **TM Finder: a prediction program for transmembrane protein segments using a combination of hydrophobicity and nonpolar phase helicity scales**. *Protein Sci* 2001, **10**(1):212-219.

21. van de Waterbeemd H, Jones BC: **Predicting oral absorption and bioavailability**. *Progress in medicinal chemistry* 2003, **41**:1-59.

22. Palm K, Luthman K, Ungell AL, Strandlund G, Beigi F, Lundahl P, Artursson P: **Evaluation of dynamic polar molecular surface area as predictor of drug absorption: comparison with other computational and experimental predictors**. *Journal of medicinal chemistry* 1998, **41**(27):5382-5392.

23. Kelder J, Grootenhuis PD, Bayada DM, Delbressine LP, Ploemen JP: **Polar molecular surface as a dominating determinant for oral absorption and brain penetration of drugs**. *Pharmaceutical research* 1999, **16**(10):1514-1519.

24. Clark DE: **Rapid calculation of polar molecular surface area and its application to the prediction of transport phenomena. 1. Prediction of intestinal absorption**. *Journal of pharmaceutical sciences* 1999, **88**(8):807-814.

25. Conradi RA, Hilgers AR, Ho NF, Burton PS: **The influence of peptide structure on transport across Caco-2 cells. II. Peptide bond modification which results in improved permeability**. *Pharmaceutical research* 1992, **9**(3):435-439.

26. Lipinski CA, Lombardo F, Dominy BW, Feeney PJ: **Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings**. *Advanced drug delivery reviews* 2001, **46**(1-3):3-26.

27. Sha'afi RI, Gary-Bobo CM, Solomon AK: **Permeability of red cell membranes to small hydrophilic and lipophilic solutes**. *J Gen Physiol* 1971, **58**(3):238-258.

28. Raevsky OA, Schaper K: **Quantitative estimation of hydrogen bond contribution to permeability and absorption processes of some chemicals and drugs**. *Eur J Med Chem* 1998, **33**:799-807.

29. Norinder U, Osterberg T: **The applicability of computational chemistry in the evaluation and prediction of drug transport properties**. *Perspectives in Drug Discovery and Design* 2000, **19**:1-18.

30. Lipinski CA: **Drug-like properties and the causes of poor solubility and poor permeability**. *J Pharmacol Toxicol Methods* 2000, **44**(1):235-249.

31. Pajouhesh H, Lenz GR: **Medicinal chemical properties of successful central nervous system drugs**. *NeuroRx* 2005, **2**(4):541-553.

32. Conradi RA, Hilgers AR, Burton PS, Hester JB: **Epithelial cell permeability of a series of peptidic HIV protease inhibitors: aminoterminal substituent effects**. *Journal of drug targeting* 1994, **2**(2):167-171.

33. Rezai T, Bock JE, Zhou MV, Kalyanaraman C, Lokey RS, Jacobson MP: **Conformational flexibility, internal hydrogen bonding, and passive membrane permeability: successful in silico prediction of the relative permeabilities of cyclic peptides**. *Journal of the American Chemical Society* 2006, **128**(43):14073-14080.

34. Rezai T, Yu B, Millhauser GL, Jacobson MP, Lokey RS: **Testing the conformational hypothesis of passive membrane permeability using synthetic cyclic peptide diastereomers**. *Journal of the American Chemical Society* 2006, **128**(8):2510-2511.

35. Roseman MA: **Hydrophobicity of the peptide C=O...H-N hydrogen-bonded group**. *Journal of molecular biology* 1988, **201**(3):621-623.

36. Wright LL, Painter GR: **Role of desolvation energy in the nonfacilitated membrane permeability of dideoxyribose analogs of thymidine**. *Molecular pharmacology* 1992, **41**(5):957-962.

37. el Tayar N, Mark AE, Vallat P, Brunne RM, Testa B, van Gunsteren WF: **Solvent-dependent conformation and hydrogen-bonding capacity of cyclosporin A: evidence from partition coefficients and molecular dynamics simulations**. *Journal of medicinal chemistry* 1993, **36**(24):3757-3764.

38. Pagliara A, Reist M, Geinoz S, Carrupt PA, Testa B: **Evaluation and prediction of drug permeation**. *The Journal of pharmacy and pharmacology* 1999, **51**(12):1339-1357.

39. Veber DF, Johnson SR, Cheng HY, Smith BR, Ward KW, Kopple KD: **Molecular properties that influence the oral bioavailability of drug candidates**. *Journal of medicinal chemistry* 2002, **45**(12):2615-2623.

40. van De Waterbeemd H, Smith DA, Beaumont K, Walker DK: **Property-based design: optimization of drug absorption and pharmacokinetics**. *Journal of medicinal chemistry* 2001, **44**(9):1313-1333.

CHAPTER SIX
Perspectives

6.1 ENGINEERING TOOLS

Numerous biomedical engineering technologies that interface molecular biology, protein chemistry, analytical chemistry, computer science and statistics were applied collectively in the search for protein disease markers in Chapters 3 and 4. In these two case studies, mass spectrometry was the main enabling technology for proteomics. Even though it has long been used in analytical chemistry for small molecule analysis, mass spectrometry only became the dominant platform for the study of biomolecules upon the development of soft ionization techniques such as MALDI and ESI. These techniques facilitate analysis of the intact biomolecules by allowing them to be converted to ions without inducing fragmentation. Innovative engineering advances also led to both the development of array chips with chemically derivatized surfaces to selectively capture a subset of proteins for high-throughput analysis at the molecular level and the development of an albumin enrichment platform based on Cibachron blue for the selective affinity capture of albumin. In our studies, we coupled these two proteome simplification technologies to the high-resolution, high-mass accuracy prOTOF mass spectrometer via a custom made adapter capable of handling array chips to create a powerful platform for high-throughput protein profiling of diseases. Once mass peaks representative of the sample groups were obtained, computational platforms developed by computer scientists were employed for data analysis. Bioinformatics engineering included the development of a novel data analysis approach in-house using Perl scripts, the implementation of a more elaborate logistic regression protocol encoded in the SAS program and the application of four unique statistical algorithms to the analysis of the same mass spectral data set for the discovery of differential peaks with high discriminatory power. These peaks were then enriched and the proteins identified by mass spectrometry based on tandem mass sequencing and protein database search using numerous database search algorithms that were developed and made available over the years.

6.2 CONCLUSIONS AND PERSPECTIVES

There is great interest in the discovery of new protein markers for a variety of diseases to aid in diagnosis, prognosis and evaluation of therapeutic responsiveness. Mass spectrometry has proven to be the backbone technology of proteomics biomarker discovery, conferring the ability to interrogate thousands of proteins simultaneously in a high-throughput and facile manner.

Differential protein pattern profiling has the potential to be instated as a non-invasive, rapid test that provides diagnostic utility to assist in the clinical decision-making process. Independent of the identity of the peptides or proteins, the intensities of the *m/z* peaks form the discriminator and may be clinically applicable before their identities are discerned. This approach does not require the lengthy development and validation of antibody reagents for immunoassay-based systems, such as costly antibody production and purification, and the subsequent tests for heterophillic antibody interferences. Moreover, the multimarker profiles have demonstrated both high diagnostic sensitivities and specificities, a desired trait unattainable by most single disease biomarkers of clinical currency.

However, as a technology still growing out of its infancy, numerous notable hurdles that constitute the major roadblock to its clinical utility (as discussed in Chapter 2) remain to be overcome. The limited sensitivity of current mass spectrometers has been the main impediment to the coverage depth of the proteome (Fig. 6.1). Fractionation strategies could extend this by two orders of magnitude but the compromise lies in achieving a balance between depth of coverage and throughput.

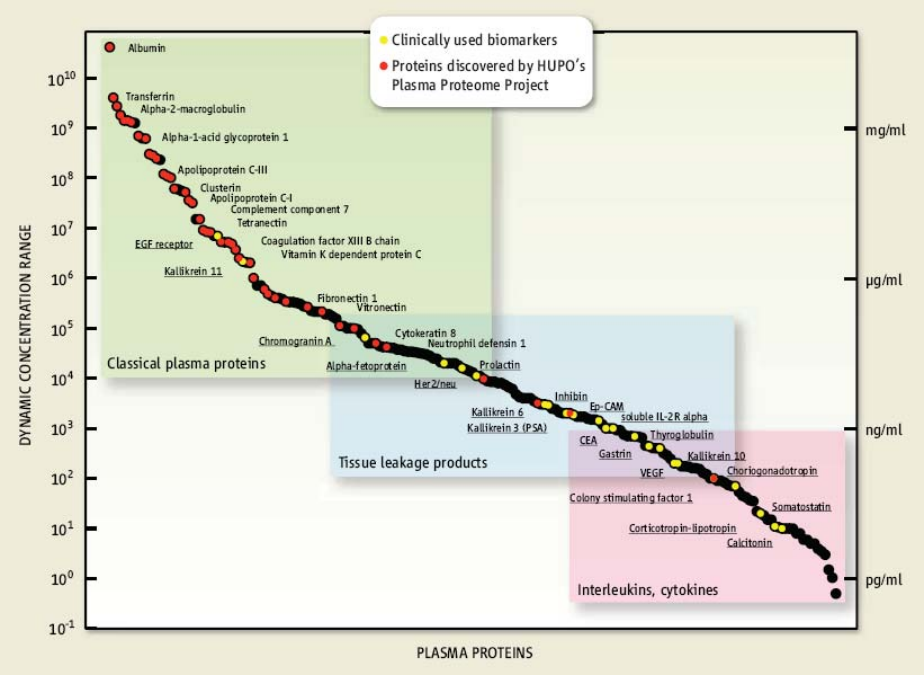


Figure 6.1 **Sensitivity and proteome depth of coverage**. The difference in blood protein abundance spans 10 orders of magnitude, resulting in the disparity between detectability (proteins identified to date, red) and applicability (biomarkers in current use, yellow). [1]

The marriage of surface retentate chemistry and mass spectrometry in SELDI TOF MS is a powerful technology that permits on-chip fractionation of a complex biological sample prior to mass spectrometry readout. Even though the process reduces the analysis to just a subset of the proteome that is dependent on the capture surface chemistry used, the depth of proteome coverage can easily be extended by a combination of different chip surfaces in the analysis of the same sample [2]. This is feasible due to the high-throughput nature of the platform.

Current SELDI approaches have come under immense criticisms for its lack of reproducibility as well as biological and technical biases. These challenges have in turn

motivated gradual improvements to the technology and more still need to be undertaken, especially in the experimental design, if it is to see eventual clinical utility.

In Chapter 2, we described a workflow aimed at addressing and controlling the critical issues of reproducibility, variability (biological, analytical and stochastic) and false positives, where the entire process from sample processing to data generation and the eventual data analysis was monitored. The method described represents a rigorous experimental design that serves to address the main challenges listed in Chapter 1 for mass spectrometry-based proteomics biomarker discovery, as well as the often overlooked sources of variability specific to SELDI. We sought to increase sensitivity by matching the dynamic range of our samples to the mass spectrometer, either through concentration of analytes in proximal fluids or proteome simplification. We improved on resolution and mass accuracy by adopting a high performance mass spectrometer to generate more peaks in the spectrum. We demonstrated high reproducibility by controlling for analytical variations (automation, same batch of array chips, same operator and minimal mass drift). The experimental design also controls for biological variations by analyzing multiple samples representative of the target population and for analytical variations by running replicates. Furthermore, we implemented a novel robust statistical data analysis approach to reduce false positives from overfitting. In addition to selecting differential peaks, our data analysis approach narrows the list to the more discriminatory ones as candidate markers for identification and verification efforts.

The robustness and utility of this optimized platform was evaluated and demonstrated in two independent autoimmune disease studies. In Chapter 3, a case study of multiple sclerosis (MS) was undertaken with sixty representative clinical CSF samples from both MS and control groups. A differential protein profile with good discriminatory power (AUC= 0.76) was obtained for the classification of MS in general. In addition, a differential peak was found to be preferentially higher in the secondary progressive stage in the subgroup comparison within MS.

Further investigation led to the identification and verification of Complement C3 as a differential marker in the clinical samples using Western blot.

In Chapter 4, proteomics biomarker discovery was conducted using the same workflow with thirty serum samples representative of narcoleptic and non-narcoleptic patients from the Center for Narcolepsy at Stanford University. The non-narcoleptic group includes those who share the same susceptibility genetic background as the narcoleptic patients but do not display any symptoms of narcolepsy. In this pilot study, analysis was only performed on the bound species of albumin because of their known diagnostic value and also as a venue to simplify the serum proteome. Differential protein profiles consisting of robust biomarkers were obtained with great discriminatory power for each group comparison. In particular, a differential peak found in the comparison between narcoleptic and non-narcoleptic patients in the presence of the HLA DQB1*0602 susceptibility gene was successfully identified as a fragment of the bikunin protein. Subsequent verification using Western blot analysis between the two groups confirmed its preferential higher level in the narcolepsy group.

The successful identification of differentially present proteins in these two studies demonstrate the robustness of this workflow that is applicable to diseases as complex as MS and as uncommon as narcolepsy. Given these proteins were found to be elevated in the disease group and not the control makes them ideal biomarkers as they are disease-specific. In addition, the fact that the biomarkers Complement C3 and bikunin have been implicated in the inflammatory processes that could be responsible for the pathogenesis of these two autoimmune diseases stresses the importance of sound experimental design that incorporate disease specific knowledge in the discovery process through the selection of appropriate samples. It is especially encouraging given our discovery that Complement C3 could contribute to disease severity in MS correlates extremely well with its role in maximal disease progression in EAE, the animal model of MS.

The two case studies presented in this dissertation represent the exploratory phase of biomarker discovery, where disease marker discovery and preliminary verification efforts were

performed within the limited sample size available for each study. They form the basis for the subsequent validation phase where these known protein markers can be assessed for diagnostic potential in a much larger sample group (Fig. 6.2). It is foreseeable for the identified potential protein markers to be implemented in an ELISA-like assay to analyze more patient samples in a high-throughput manner to determine its diagnostic utility. Once their diagnostic potential is established, they can be developed into clinical assays to facilitate disease diagnosis, prognosis and therapeutic efficacy evaluation as well as serve as drug targets. This is especially true in the case of Complement C3 which our data suggest could be adopted as a stage-specific marker to differentiate between the more drug-treatable, relapsing-remitting stage and the chronic, secondary progressive stage within MS itself.

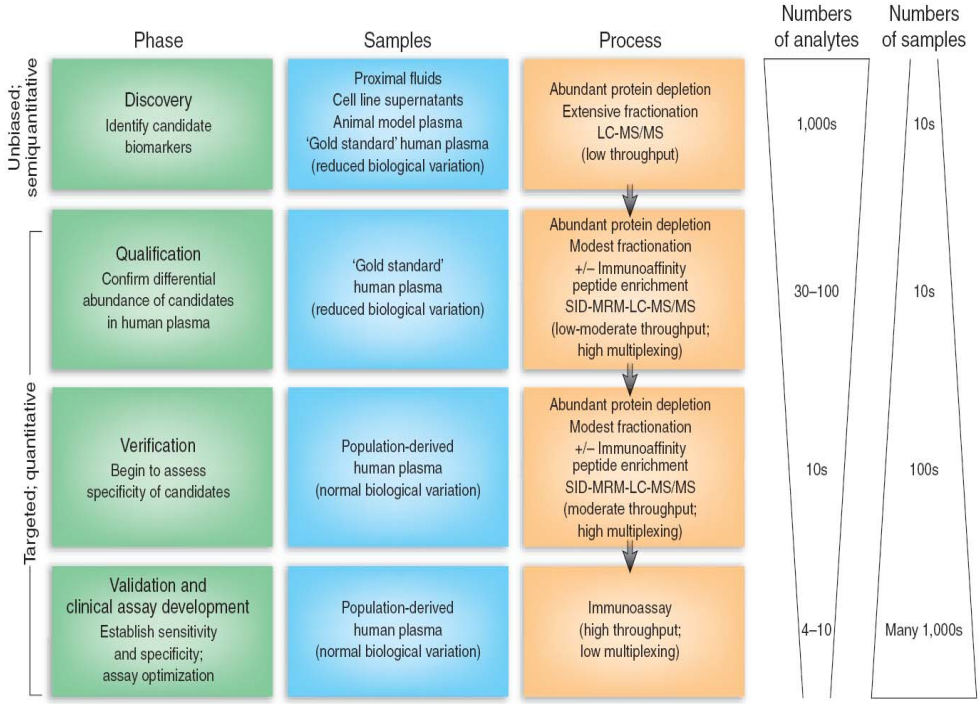


Figure 6.2 **Phases of biomarker discovery and validation**. The gradual progression from the initial, unbiased discovery phase to the hypothesis-driven validation stage sees the reduction in the number of analytes sampled and an increase in the number of samples that are more representative of the target population. [3]

As it turns out, the high fidelity of the output spectra garnered from the optimized workflow allows us to revisit the notion of disease classification based solely on the identity-free diagnostic model. Previous doubts about this approach lied on its lack of reproducibility and biological inference to the disease process. In our case studies, we showed that reproducible protein profiles can be reliably generated when a high performance mass spectrometer is adopted. We see minimal mass drift among the peaks across spectra with a mass accuracy of <10 ppm while the signal intensity has a CV of 5% -10% in agreement with others, which is well within the CVs for established markers used in clinical diagnostics estimated to be in the range of 1.5% - 10%. It is foreseeable that a reevaluation of this approach using a high performance mass spectrometer to demonstrate minimal interlaboratory variation will confirm and justify the utility of the differences between MS profiles of the disease and control specimens to generate a diagnostic model in the clinics. The need to unravel the peak identity before clinical adaptation is unfounded given the historical origin of PSA and CA-125.

Until platform reproducibility can be demonstrated across sites in an initiative similar to the one with low performance mass spectrometers [4], this technology will likely remain a useful front-end discovery tool for biomarkers. Even if mass peak reproducibility can be guaranteed, there still lies the problem of detecting low abundance proteins. A recent reported approach that can bridge the span of protein concentration in complex biological samples and the limited dynamic range of current proteomics detection methods involves the use of combinatorial ligand libraries [5, 6]. The reasoning is that given the diverse library comprises of all possible ligands for the binding of both high and low abundance proteins, even though the diversity of the protein mixture remains unchanged, the chances of low abundance proteins finding a binding partner and enriched for detection are greatly enhanced. This is based on an established premise in combinatorial chemistry stating in order to probe a diverse biological space, an equally diverse chemical space is needed. Combinatorial libraries have proven successful in the analysis of the previously unseen proteome in human urine, serum, and platelets lysate [7-9].

Another persistent challenge lies in controlling the preanalytical variability during sample procurement as proteins are sensitive to storage, handling, and processing conditions (Chapter 2). Therefore, protein profiling may be more reliable if it were to focus on proteins that are less prone to these biases, such as antibodies. This is particularly appealing in the study of autoimmune diseases whose trademark is the presence of autoantibodies. Compared to traditional protein biomarkers, this subset of the proteome carries the advantages of improved sample stability and non-fluctuating levels in sera, making them highly effective biomarkers. The multiplex analyses of autoantibodies for disease 'signatures' that can confer the same level of sensitivity and specificity of a mass spectrometry profile for diagnosis have been reported [10-12]. The downside of this approach, however, is limitation to the existing known autoantigens. Combinatorial chemistry may again present the solution to this bottleneck where libraries of small molecules can be screened for novel binders to autoantibodies as a replacement for autoantigens. Admittedly, the binding affinity of antibodies to their small molecule binder might be significantly lower (in the low μM range) and thus result in low specificity. However, a panel of these independent weaker binders can collectively confer the high sensitivity and specificity desired, as seen in mass spectrometry protein profiles. The class of molecules that is suited for this endeavor should preferably be immune to potential interferences from the biological constituents of biological samples, such as protease activities in blood.

Peptoids as the protease-resistant analogs of peptides [13] are ideal for the screening of the binders aforementioned. Their facile route of synthesis allows a more diverse set of functional groups to be incorporated than peptides, effectively expanding the chemical space to approximate the diversity in antibodies naturally conferred by VDJ recombination. Additionally, as presented in Chapter 5, we have shown that peptoids are, in general, more cell permeable than peptides. This makes them an attractive alternative for drug development to target protein markers in circulation and perhaps even autoreactive entities in the brain across the BBB.

Clearly, six years into the seminal report on the diagnostic application of SELDI, many technological issues remain to be resolved. There is the need to carefully define a common operating procedure essential for the validation of this technology in anticipation of its introduction into clinical practice. Efforts toward this standardization have been initiated across disciplines by organizations such as HUPO and the American Association of Clinical Chemistry. Meanwhile, ongoing investigations are slowly but surely providing innovative solutions to address the obstacles that currently define SELDI. Critical experimental design and adoption of a workflow that primarily aims at reducing bias and securing reproducibility such as the one presented in this dissertation contribute to the betterment of this promising technology. Now is indeed an exciting time to be part of a global collaborative effort to fine tune a maturing technology to achieve its full potential in revolutionizing biomarker discovery for its multifarious roles of disease risk determination, early detection and diagnosis, disease staging, prognosis, therapeutic options evaluation, and monitoring of responsiveness to treatments, all to realize the simple goal of improving health and prolonging life.

6.3 BIBLIOGRAPHY

1. Service RF: **Proteomics. Will biomarkers take off at last?** *Science (New York, NY* 2008, **321**(5897):1760.
2. Guerreiro N, Gomez-Mancilla B, Charmont S: **Optimization and evaluation of surface-enhanced laser-desorption/ionization time-of-flight mass spectrometry for protein profiling of cerebrospinal fluid**. *Proteome science* 2006, **4**:7.
3. Rifai N, Gillette MA, Carr SA: **Protein biomarker discovery and validation: the long and uncertain path to clinical utility**. *Nature biotechnology* 2006, **24**(8):971-983.
4. Semmes OJ, Feng Z, Adam BL, Banez LL, Bigbee WL, Campos D, Cazares LH, Chan DW, Grizzle WE, Izbicka E *et al*: **Evaluation of serum protein profiling by surface-enhanced laser desorption/ionization time-of-flight mass spectrometry for the detection of prostate cancer: I. Assessment of platform reproducibility**. *Clinical chemistry* 2005, **51**(1):102-112.
5. Righetti PG, Boschetti E, Lomas L, Citterio A: **Protein Equalizer Technology : the quest for a "democratic proteome"**. *Proteomics* 2006, **6**(14):3980-3992.
6. Boschetti E, Giorgio Righetti P: **Hexapeptide combinatorial ligand libraries: the march for the detection of the low-abundance proteome continues**. *BioTechniques* 2008, **44**(5):663-665.
7. Castagna A, Cecconi D, Sennels L, Rappsilber J, Guerrier L, Fortis F, Boschetti E, Lomas L, Righetti PG: **Exploring the hidden human urinary proteome via ligand library beads**. *Journal of proteome research* 2005, **4**(6):1917-1930.
8. Sennels L, Salek M, Lomas L, Boschetti E, Righetti PG, Rappsilber J: **Proteomic analysis of human blood serum using peptide library beads**. *Journal of proteome research* 2007, **6**(10):4055-4062.

9. Guerrier L, Claverol S, Fortis F, Rinalducci S, Timperio AM, Antonioli P, Jandrot-Perrus M, Boschetti E, Righetti PG: **Exploring the platelet proteome via combinatorial, hexapeptide ligand libraries**. *Journal of proteome research* 2007, **6**(11):4290-4303.

10. Hueber W, Robinson WH: **Proteomic biomarkers for autoimmune disease**. *Proteomics* 2006, **6**(14):4100-4105.

11. Villalta D, Tozzoli R, Tonutti E, Bizzaro N: **The laboratory approach to the diagnosis of autoimmune diseases: is it time to change?** *Autoimmunity reviews* 2007, **6**(6):359-365.

12. Balboni I, Chan SM, Kattah M, Tenenbaum JD, Butte AJ, Utz PJ: **Multiplexed protein array platforms for analysis of autoimmune diseases**. *Annual review of immunology* 2006, **24**:391-418.

13. Wang Y, Lin H, Tullman R, Jewell CF, Jr., Weetall ML, Tse FL: **Absorption and disposition of a tripeptoid and a tetrapeptide in the rat**. *Biopharmaceutics & drug disposition* 1999, **20**(2):69-75.

APPENDIX A
Clinical samples in Case Study I: Multiple Sclerosis (Chapter 3)

Clinical CSF samples used in the study with patient group classification, diagnosis, gender, and age: MS = multiple sclerosis, RRMS = relapsing-remitting multiple sclerosis, SPMS = secondary progressive multiple sclerosis, OND = other neurological diseases, PD = Parkinson's Disease.

Index	Sample	Group	Diagnosis	Gender	Age
1	M558	MS	RRMS	F	29
2	M875	MS	RRMS	F	22
3	M584-3	MS	RRMS	F	43
4	M125	MS	RRMS	F	32
5	M354	MS	RRMS	F	42
6	M522-1	MS	RRMS	F	34
7	M376-1	MS	RRMS	F	57
8	M918-1	MS	RRMS	M	37
9	M465-1	MS	RRMS	F	31
10	M584-2	MS	RRMS	F	41
11	M125-2	MS	RRMS	F	33
12	M746	MS	RRMS	F	42

Index	Sample	Group	Diagnosis	Gender	Age
13	M818	MS	RRMS	F	60
14	M927	MS	RRMS	F	30
15	6495	MS	SPMS	M	25
16	6519	MS	SPMS	M	44
17	6613	MS	SPMS	M	43
18	6592	MS	SPMS	M	49
19	6620	MS	SPMS	M	52
20	6721	MS	SPMS	M	30
21	6807	MS	SPMS	F	42
22	6837	MS	SPMS	M	37
23	6963	MS	SPMS	M	54
24	7250	MS	SPMS	F	67
25	7464	MS	SPMS	M	37
26	7509	MS	SPMS	M	38
27	9593	MS	SPMS	M	38
28	9603	MS	SPMS	M	39
29	10238	MS	SPMS	F	36
30	11621	MS	SPMS	F	28

Index	Sample	Group	Diagnosis	Gender	Age
31	11620	MS	SPMS	F	48
32	11757	MS	SPMS	M	54
33	6935	MS	SPMS	M	55
34	6876	MS	SPMS	M	66
35	M142	Non-MS	OND	F	30
36	M636	Non-MS	OND	M	60
37	M758	Non-MS	OND	F	43
38	12265	Non-MS	Prostate Cancer	M	79
39	12286	Non-MS	Prostate Cancer	M	77
40	12281	Non-MS	Urinary/Bladder Cancer	M	73
41	12292	Non-MS	Congestive Heart Failure	F	91
42	12300	Non-MS	Congestive Heart Failure	M	86
43	10558	Non-MS	Headache	M	44
44	11598	Non-MS	Headache	M	41
45	7682	Non-MS	Headache	M	38
46	6070	Non-MS	Seizure	M	47
47	6232	Non-MS	Seizure	M	29
48	6247	Non-MS	Seizure	M	49

Index	Sample	Group	Diagnosis	Gender	Age
49	6289	Non-MS	Seizure	M	26
50	6373	Non-MS	Seizure	M	64
51	11884	Non-MS	PD	F	80
52	11891	Non-MS	PD	M	75
53	12089	Non-MS	PD	M	70
54	12099	Non-MS	PD	M	84
55	12133	Non-MS	PD	M	80
56	8011	Non-MS	Cerebellar Tumor	M	44
57	9768	Non-MS	Glioblastoma	M	69
58	12893	Non-MS	Glioblastoma	M	52
59	12872	Non-MS	Brain Tumor	F	92
60	12112	Non-MS	Brain Tumor	M	48

APPENDIX B
Sequence information data of highly permeable OxDex-conjugated molecules for high-throughput study (Chapter 5)

Sequence information of highly permeable OxDex-conjugated peptoid tetramers. Highly permeable peptoids have a permeability ratio (PR) that is at least 2σ above the average permeability ratio of all peptoids in the high-throughput study. The frequency of each monomer is indicated for each peptoid.

PEPTOIDS	PR	Molecular weight	Sequence Composition				
			Nmba	Nleu	Nlys	Ngly	Nser
4merTOID PL07B10	0.0213	1052	0	1	1	1	1
4merTOID PL07B11	0.0213	1085	1	1	0	1	1
4merTOID PL09B01	0.0221	1039	0	1	0	2	1
4merTOID PL09E03	0.0206	1081	0	0	2	2	0
4merTOID PL09H11	0.0229	1050	0	2	1	0	1
4merTOID PL10C03	0.0218	1053	0	1	0	3	0
4merTOID PL10G03	0.0228	1037	0	2	0	1	1
4merTOID PL10H03	0.0244	1119	2	0	0	0	2
4merTOID PL10G07	0.0216	1064	0	2	1	1	0
4merTOID PL10D08	0.0206	1054	0	0	1	2	1
4merTOID PL10E10	0.0223	1077	0	2	2	0	0
4merTOID PL10H10	0.0232	1038	0	1	1	0	2
Average	0.0221	1062					

Sequence information of highly permeable OxDex-conjugated peptide tetramers. Highly permeable peptides have a permeability ratio (PR) that is at least 2σ above the average permeability ratio all peptides in the high-throughput study. The frequency of each monomer is indicated for each peptide.

PEPTIDES	PR	Molecular weight	Sequence Composition				
			Phe	Val	Lys	Asp	Ser
4merTIDE PL05F06	0.0126	1071	1	1	0	2	0
4merTIDE PL05F07	0.0118	1053	0	0	2	1	1
4merTIDE PL05H11	0.0156	1097	1	1	2	0	0
4merTIDE PL06F11	0.0115	1075	2	1	0	0	1
4merTIDE PL06G11	0.0124	1063	2	0	0	0	2
4merTIDE PL07B02	0.0141	1085	1	0	2	0	1
4merTIDE PL08H03	0.0171	1084	1	1	1	1	0
4merTIDE PL08H04	0.0170	996	0	1	1	0	2
4merTIDE PL08H05	0.0117	1145	2	0	2	0	0
4merTIDE PL08D07	0.0131	1024	0	1	1	1	1
4merTIDE PL08H07	0.0119	1132	2	0	1	1	0
4merTIDE PL08H10	0.0120	1072	1	0	1	1	1
Average	0.0134	1074					