HIGH THROUGHPUT PROTEOMIC STUDIES USING FOURIER TRANSFORM ION CYCLOTRON RESONANCE MASS SPECTROMETRY

by

STACEY A. BOLTZ

(Under the Direction of I. JONATHAN AMSTER)

ABSTRACT

With the advent of soft ionization techniques, namely matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI), mass spectrometry has emerged as the method of choice for proteomic analysis. Traditionally, two-dimensional gel electrophoresis (2D-GE) followed by mass spectrometry has been utilized to identify proteins. Due to numerous limitations of 2D-GE, high-throughput alternatives have been sought out. Shotgun proteomics allows for the analysis of an entire proteome simultaneously through batch digestion of whole-cell lysates. Proteins can be identified from the resulting complex mixture from accurate mass measurements of their constituent peptides. Fourier transform mass spectrometry (FTMS) is capable of achieving the part-per-million mass accuracy and ultra-high mass resolution needed for these measurements. This thesis details a high-throughput proteomic method using LC-MALDI-FTMS to analyze a protein standard and a cell lysate. Additionally, a novel technique called mass defect labeling is described as a way to increase the specificity of accurate mass measurements.

INDEX WORDS: Mass spectrometry, FTMS, Proteomics, MALDI, ESI, Accurate mass measurement, Mass defect labeling

HIGH THROUGHPUT PROTEOMIC STUDIES USING FOURIER TRANSFORM ION CYCLOTRON RESONANCE MASS SPECTROMETRY

by

STACEY A. BOLTZ

B.S., The University of Richmond, 2000

A Thesis Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

MASTER OF SCIENCE

ATHENS, GEORGIA

2003

HIGH THROUGHPUT PROTEOMIC STUDIES USING FOURIER TRANSFORM ION CYCLOTRON RESONANCE MASS SPECTROMETRY

by

STACEY A. BOLTZ

Major Professor:	I. Jonathan Amster
Committee:	Ron Orlando John Stickney

Electronic Version Approved:

Maureen Grasso
Dean of the Graduate School
The University of Georgia
May 2003

To my friends and family… You may not have quite understood what I have immersed myself in the past three years, but you have supported me regardless

ACKNOWLEDGEMENTS

I would like to thank my colleagues in the chemistry department, especially my fellow group members, Julia Swancy, Bryan Parks, and Jeremy Wolff. While their ideas and assistance in the lab has been invaluable, their sense of humor and willingness to converse about all random non-chemistry related topics is what has made my experience at UGA memorable.

I would also like to thank the people at Bruker Daltonics for their financial and technical support, especially former Amster group member Mike Easterling. My gratitude is also extended to Dr. Barny Whitman for his financial assistance and interest in proteomics. My appreciation also goes out to his post-doc Iris Porat for providing our lab with the proteins of *M. maripaludis* for analysis. My thanks is additionally extended to Dr. Robbie Phillips for being patient with my complete lack of organic chemistry knowledge, to Melanie Brown for synthesizing the mass defect label, and to professors Ron Orlando and John Stickney for taking time out of their busy schedules to serve on my committee. Last, but certainly not least, I owe a debt of gratitude to my advisor, Professor Jon Amster, for his understanding, support, and seemingly never-ending string of ideas…

TABLE OF CONTENTS

Page

ACKNOWLEDGEMENTS v

LIST OF TABLES viii

LIST OF FIGURES ix

CHAPTER

1 LITERATURE REVIEW 1

Biological Mass Spectrometry 2

Proteomics 17

References 35

2 EXPERIMENTAL 42

Instrumentation 43

Sample Preparation 54

References 59

3 A LC-MALDI-FTMS BASED METHOD FOR HIGH-THROUGHPUT PROTEOMICS 61

Introduction 62

Experimental 67

Results and Discussion 72

Conclusions 85

References 89

4 INCREASING SPECIFICITY OF ACCURATE MASS MEASUREMENTS USING MASS DEFECT LABELING OF PEPTIDES 92

Introduction 93

Experimental 101

Results and Discussion 102

Conclusions 106

References 110

5 CONCLUSIONS 111

References 114

APPENDIX A 115

APPENDIX B 118

LIST OF TABLES

Table Page

1.1 The advantages and disadvantages of matrix-assisted laser desorption ionization (MALDI) 7

1.2 The advantages and disadvantages of electrospray ionization (ESI) 12

3.1 Analysis of BSA digest illustrating the high mass accuracy capability of LC-intermediate pressure MALDI-FTMS 73

3.2 The number of proteins of a given molecular weight and sequence coverage range identified from 1462 randomly generated tryptic peptide masses of *Methanococcus maripaludis* 86

3.3 The number of proteins of a given molecular weight and sequence coverage range identified from 1462 experimentally obtained tryptic peptide masses of a sub-20 kDa *Methanococcus maripaludis* digest 87

LIST OF FIGURES

Figure Page

1.1 A schematic of a matrix-assisted laser desorption ionization (MALDI) source........4

1.2 A schematic of an electrospray ionization (ESI) source..8

1.3 A diagram of ion formation from an electrospray ionization source 10

1.4 The relationship between cyclotron motion and the Lorentz force, F, for an ion of velocity, ν, in a magnetic field, B 13

1.5 An illustration of ion excitation and detection in a Fourier transform mass spectrometry (FTMS) analyzer cell...16

1.6 An illustration of the relationship between transient duration and mass resolution in Fourier transform mass spectrometry (FTMS)..18

1.7 A schematic of protein identification by 2-dimensional gel electrophoresis (2D-GE) and tandem mass spectrometry (MS/MS)..21

1.8 A diagram of search specificity for *Escherichia coli* proteins identified from its tryptic peptides measured at 1, 10, and 100 ppm mass accuracy...........................25

1.9 A diagram of search specificity for *Escherichia coli* proteins identified from its cysteine-containing tryptic peptides measured at 1, 10, and 100 ppm mass accuracy...27

1.10 An illustration of isotope-coded affinity tag (ICAT) reagents for protein quantitation. 30

1.11 An overview of quantitative isotope-coded affinity tag (ICAT) analysis...............31

2.1 A photograph of a Micro-Tech Ultra Plus II MD capillary HPLC system 44

2.2 A photograph of a reverse-phase capillary HPLC setup 45

2.3 A photograph of a Gilson 215 single-channel autosampler. 47

2.4 A photograph showing co-deposition of LC fractions and matrix on a 49-Spot MALDI target...49

2.5 A schematic of a Bruker FTMS-130 intermediate pressure MALDI source 50

2.6 A photograph of a FTMS-130 intermediate pressure MALDI source coupled to a 7.0 tesla Bruker BioApex Fourier transform mass spectrometer 52

2.7 A schematic of a 4.7 tesla MALDI Fourier transform mass spectrometer.............53

3.1 An illustration of a Bio-Rad 491 preparative electrophoresis cell 69

3.2 A photograph of a SDS-PAGE of the low molecular weight protein fractions of *Methanococcus maripaludis*...70

3.3 A mass spectrum of *Methanococcus maripaludis* peptides illustrating the high mass resolution capability of LC-intermediate pressure MALDI-FTMS 75

3.4 A mass spectrum of a BSA digest showing metastable losses using vacuum MALDI-FTMS 76

3.5 A mass spectrum of a BSA digest demonstrating suppression of metastable losses using LC-intermediate pressure MALDI-FTMS...77

3.6 A two-dimensional contour plot and a horizontal mass spectral projection of a Sub-20 kDa *Methanococcus maripaludis* protein tryptic digest produced by LC-intermediate pressure MALDI-FTMS 79

3.7 A combined mass spectrum of all LC fractions of a sub-20 kDa *Methanococcus maripaludis* protein tryptic digest analyzed by intermediate pressure MALDI-FTMS 80

3.8 Mass spectra of LC Fraction 2 of a Sub-20 kDa *Methanococcus maripaludis* protein tryptic digest analyzed by intermediate pressure MALDI-FTMS 82

3.9 A histogram of protein molecular weight vs. sequence coverage distribution from 1462 randomly generated tryptic peptide masses of *Methanococcus maripaludis* 85

3.10 A histogram of protein molecular weight vs. sequence coverage distribution from 1462 experimentally obtained tryptic peptide masses of a sub-20 kDa *Methanococcus maripaludis* digest......86

4.1 A histogram demonstrating the narrow distribution of theoretical tryptic peptide masses from *Methanococcus maripaludis* 95

4.2 An illustration of the mass defect labeling reagent N-(2,4-dibromophenyl) maleimide 96

4.3 A schematic showing the addition of N-ethylmaleimide to the free sulfhydryl of a cysteine side-chain 97

4.4 A histogram demonstrating the mass defect distribution of *Methanococcus maripaludis* tryptic peptides......99

4.4 The modeled isotopic pattern of laminin peptide in a mass spectrum (A); and MALDI-FTICR mass spectrum of laminin peptide (B)......104

4.5 The modeled isotopic pattern of laminin peptide labeled with mass defect reagent in a mass spectrum (A); and MALDI-FTICR mass spectrum of laminin peptide labeled with mass defect reagent (B) 105

4.7 A MALDI-FTICR mass spectrum of laminin peptide labeled with 1-fold molar excess mass defect reagent 107

4.8 A MALDI-FTICR mass spectrum of laminin peptide labeled with 5-fold molar excess of mass defect reagent 108

CHAPTER 1

LITERATURE REVIEW

BIOLOGICAL MASS SPECTROMETRY

Mass spectrometry (MS) is a powerful analytical technique to efficiently measure molecular masses. MS measures the mass-to-charge ratio (*m/z*) of gas phase ions. With broad mass range capabilities (up to 1 million Daltons), low sample requirements (attomoles to femtomoles), and short acquisition times (μseconds), mass spectrometry is an ideal method for the analysis of biomolecules.[1] In addition, the mass accuracy and resolution of even the simplest mass spectrometer is orders of magnitude greater than alternative mass measurement methods such as gel electrophoresis or gel filtration.[2]

MS requires gas phase ions for measurement of *m/z*. MS cannot analyze neutrals because mass dependent motion that determines *m/z* necessitates an electrical charge on the molecule. All mass spectrometers consist of three main components – an ionization source, a mass analyzer, and an ion collector or detector. There are many different types of each component resulting in a variety of mass spectrometers with various capabilities and configurations depending upon the task at hand.

The ionization source is responsible for generating gas phase ions from analyte molecules. Early mass spectrometric ionization methods were not conducive to the analysis of biomolecules. Methods such as electron impact ionization (EI) and chemical ionization (CI) are not able to effectively ionize thermally labile biomolecules like proteins.[3,4] Although methods such as fast-atom bombardment (FAB), ^{252}Cf plasma desorption, and field desorption/ionization allow biomolecules to be analyzed by MS, it was the advent of two ionization methods in the 1980s – matrix-assisted laser desorption ionization (MALDI) and electrospray ionization (ESI) – that led to an explosive growth in biological applications of mass spectrometry.[5-7] The desired effect of these "soft" ionization methods is the formation of an intact molecular ion. The gentleness of MALDI and ESI imparts enough energy to ionize and vaporize large, nonvolatile biomolecules without inducing significant fragmentation.

Matrix-Assisted Laser Desorption Ionization. MALDI originated in 1988 from two independent research groups headed by Franz Hillenkamp and Koichi Tanaka.[8,9] Each research group discovered that compounds that strongly absorb radiation from a laser source can assist in the intact ionization of large biomolecules. In MALDI, ion formation is achieved by directing a pulsed laser beam at analyte molecules embedded in a matrix (Figure 1.1). The absorption of the laser radiation energy by the matrix leads to desorption of both analyte and matrix molecules producing a plume of gaseous ions and molecules.[10] The matrix acts as an energy buffer for the fragile biomolecules. It absorbs the majority of the laser radiation, causing explosive evaporation. The sample molecules are entrained in the matrix, causing them to be vaporized. Once in the gas-phase, it is believed that the matrix assists in the ionization of analyte molecules by acting as a source of protons.[1,8,11] However, the exact mechanism of the ionization process is still under investigation.[12]

Matrices used for MALDI experiments are typically small organic compounds that strongly absorb laser radiation and readily sublime. It is also necessary for the matrix to be soluble in the analyte solution. This allows the analyte molecules to incorporate within the crystal lattice of the matrix thereby reducing intermolecular forces between analyte molecules and preventing agglomeration.[13] In the first MALDI

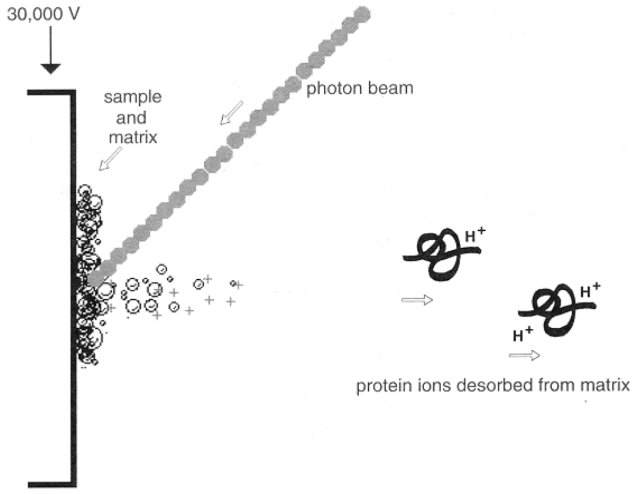

Figure 1.1 A schematic of a matrix-assisted laser desorption ionization (MALDI) source. The image was extracted from reference 10.

experiments performed by Hillenkamp et al., nicotinic acid was chosen as the matrix. While they were able to observe an ion at *m/z* 200,000 (the trimer of the protein bovine serum albumin), numerous satellite peaks due to photochemical reactions muddled the spectrum. Since then, a variety of new matrices have been employed in MALDI experiments.[11,14-22] The most common matrices for peptide and protein analysis are UV absorbing derivatives of cinnamic and benzoic acids. 3,5-dimethoxy-4-hydroxycinnamic acid (sinapinic acid), α-cyano-4-hydroxycinnamic acid (CHCA), and 2,5-dihydroxybenzoic acid (DHB) are commonly used for peptide analysis, while sinapinic acid is additionally the matrix of choice for protein analysis.[13-15]

MALDI under ultra-high vacuum causes many peptides to undergo metastable loss of one or more neutral molecules, usually water or ammonia.[23,24] Peaks resulting from these metastable losses are 17 or 18 mass units less than the monoisotopic peptide peak and have a relative abundance of 10-30%. The use of a "cool" matrix such as DHB in MALDI analysis reduces the propensity for metastable decomposition. Infrared MALDI has also been known to reduce metastable decomposition.[25,26] Ions formed by infrared MALDI have less internal energy and hence more stable than those formed by ultraviolet MALDI.[27] Another method for suppressing metastable losses is the use of a high pressure MALDI source.[28,29] Metastable decomposition is eliminated at high pressures most likely as a result of collisional thermalization of ions.[28]

Matrix-assisted laser desorption ionization is an ideal technique for the analysis of peptides and proteins. MALDI produces mainly singly charged ions allowing for easy data interpretation. Peptide mixtures produced by chemical or enzymatic digestion of proteins are often analyzed by MALDI-MS because of its unmatched simplicity. MALDI is also useful for peptide and protein analysis because it allows for rapid and simple preparation of samples, it is very sensitive, and it is relatively tolerant of impurities.[19] Contaminants such as salts, detergents, denaturants, and buffers are often present in biological sample solutions. While MALDI is tolerant of contaminants to a certain degree, high concentrations of impurities lead to poor quality mass spectra with low resolution and signal suppression.[30] Mass spectra may even become more complicated by contaminants through the presence of strong adduct ions, including the ions of sodium and potassium. Table 1.1 summarizes these and other advantages and disadvantages of the MALDI technique.

Electrospray Ionization. The origin of electrospray ionization dates back to the work of Malcolm Dole in the 1960s. Dole demonstrated the production of gas phase polymer ions.[31] Successful interfacing of an electrospray source and a mass spectrometer was first accomplished by John Fenn and coworkers using a single quadrupole mass analyzer in 1984.[32,33]

ESI produces gaseous analyte ions from a liquid solution. A sample solution is pumped through a conductive capillary maintained at a high potential (typically 2-5 kV). A fine spray of charged droplets emerges from the capillary and is electrostatically attracted to the inlet of the mass spectrometer (Figure 1.2). The droplets desolvate (sometimes with the assistance of heat and/or drying gas) as they pass from atmospheric pressure into the vacuum system of the mass spectrometer. As the solvent evaporates, the droplets decrease in size and the electric field density on the surface increases. The repulsion of mutual charges on the droplet surface eventually exceeds the surface tension causing the droplet to explode. Ions are ejected from the droplet in an outward stream

Table 1.1 The advantages of disadvantages of matrix-assisted laser desorption ionization (MALDI).

Advantages	**Disadvantages**
Practical mass range of up to 300,000 Da. Species of much greater mass have been reported. Typical sensitivity on the order of low femtomole to low picomole. Attomole sensitivity is possible. Soft ionization with little to no fragmentation observed. Tolerance of salts in millimolar concentrations. Suitable for the analysis of complex mixtures.	Matrix background, which can be a problem for compounds below a mass of 1000 Da. This background interference is highly dependent on the matrix material. Possibility of photodegradation by laser desorption ionization.

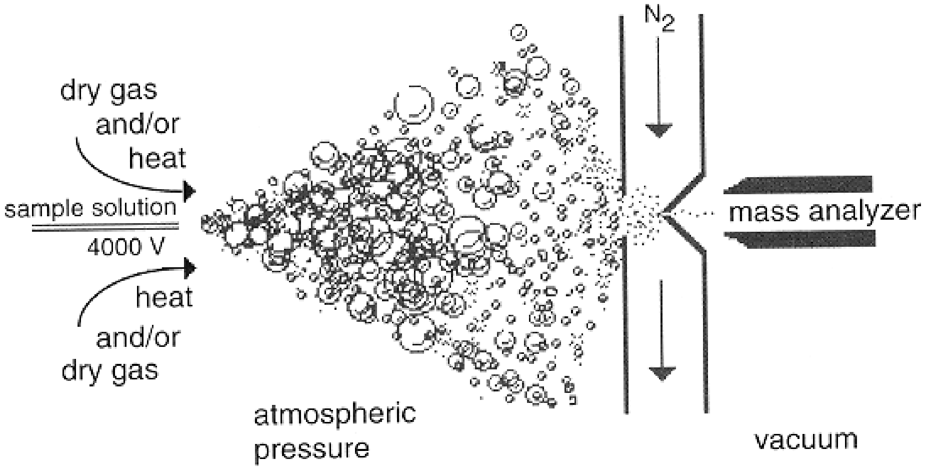

Figure 1.2 A schematic of an electrospray ionization (ESI) source. The image was extracted from reference 10.

known as a Taylor cone (Figure 1.3).[34,35] The process continues until complete desolvation occurs, and the analyte molecules are converted to gas phase ions.

Electrospray ionization made an immediate impact on protein chemistry because it produces multiply charged ions of intact peptides and proteins. This is an important feature in biological MS since mass spectrometers measure mass-to-charge ratio rather than absolute mass, thereby allowing very large proteins to be analyzed by instruments with a limited mass-to-charge range. Ions of over a million Daltons have been examined by ESI.[36] However, multiple charging makes the analysis of mixtures more complicated. Components exhibit a distribution of charge states that may overlap each other.

ESI data can be interpreted in one of three ways. The mass-to-charge spectrum can be converted into a mass spectrum by a deconvolution algorithm by Fenn.[20] A second method is to solve a simultaneous set of linear equations of mass-to-charge as a function of charge and molecular weight. These can be solved to yield the charge of each ion and the molecular weight of the analyte. Finally, with high-resolution mass spectrometers it is possible to assign charge states by measuring the isotopic peak separation.[37,38] Singly charged ions will be one mass unit apart, while doubly charged ions will be ½ mass unit apart, etc. Once charge is assigned, the molecular weight is calculated from mass-to-charge.

Equally as important to peptide and protein analysis, ESI can easily be interfaced with popular liquid-phase separation techniques. Reverse-phase high performance liquid chromatography (RP-HPLC) is often performed prior to analysis of peptide mixtures to both reduce the complexity of the mixture and to separate out possible interferents (ESI

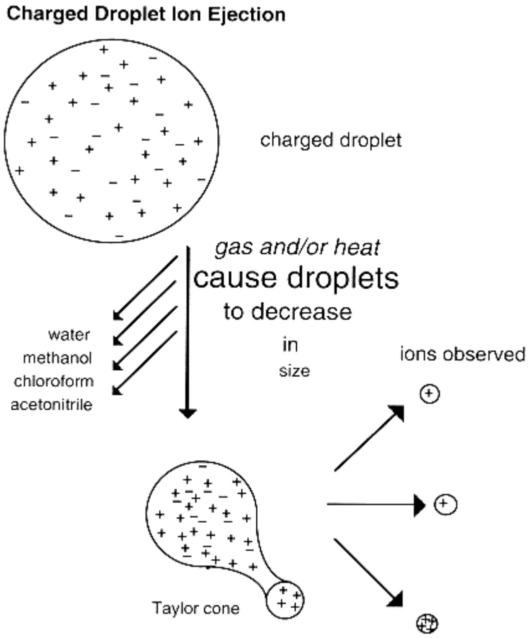

Figure 1.3 A diagram of ion formation from an electrospray ionization source. The image was extracted from reference 10.

has a low tolerance for salts). These and other advantages and disadvantages of electrospray ionization are listed in Table 1.2.

In a mass spectrometer, the mass analyzer is responsible for the separation and discrimination of ions based on their mass-to-charge ratio. Numerous mass analyzers have been easily interfaced with MALDI and ESI sources. MALDI mass spectrometers have most frequently used time-of-flight (TOF), sector, and ion-cyclotron resonance mass analyzers, while ESI sources are typically coupled with quadrupole, ion-trap, and ion-cyclotron resonance mass analyzers, although TOF is also used.[39-43] Of all mass analyzers, Fourier transform ion cyclotron resonance (FTICR) mass spectrometers achieve the highest combination of mass accuracy and resolution. The versatility and high performance capabilities of Fourier transform mass spectrometry (FTMS) make it ideal for the analysis of biomolecules.

Fourier Transform Ion Cyclotron Resonance Mass Spectrometry. First developed in the 1950's by Sommer, Thomas, and Hipple, ion cyclotron resonance (ICR) mass spectrometry utilizes the motions charged particles experience in the presence of a magnetic field to determine *m/z*.[44] In 1974, Mel Comisarow and Alan Marshall developed Fourier transform based mass spectrometry (FTMS).[45,46] When a gas-phase ion drifts into or is formed in a strong magnetic field, it experiences a force that is perpendicular to its velocity and the direction of the field. Due to this so-called Lorentz force, the ion undergoes a circular orbit known as cyclotron motion (Figure 1.4).[47] This periodic motion is defined by its cyclotron frequency, f_c, as shown by Equation 1.

$$f_c = \frac{qB}{2\pi m} \tag{1}$$

Table 1.2 The advantages and disadvantages of electrospray ionization (ESI).

Advantages	**Disadvantages**
Practical mass range up to 70,000 Da.	Low salt tolerance.
Good Sensitivity. Femtomole to low picomole sensitivity is typical.	Difficulty in cleaning overly contaminated instrument due to high sensitivity for certain compounds.
Softest ionization. Capable of observing biologically native noncovalent interactions	Low tolerance for mixtures. Simultaneous mixture analysis can be poor. The purity of the sample is important.
Easily adaptable to microbore liquid chromatography. Capable of directly analyzing LC effluent at a flow rate of one milliliter per minute.	Multiple charging, which can be confusing, especially with mixture analysis.
No matrix interference.	
Multiple charging, allowing for the analysis of high-mass ions with a relatively low m/z range instrument.	
Multiple charging, giving better mass accuracy through averaging.	

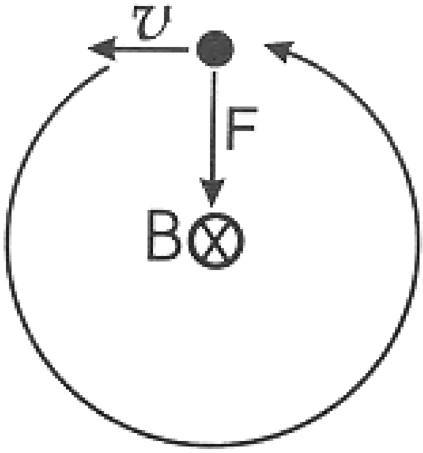

Figure 1.4 The relationship between cyclotron motion and the Lorentz force, F, for an ion of velocity, ν, in a magnetic field, B. The image was adapted from reference 47.

Cyclotron frequency is determined by three physical parameters – the charge on the ion, q, the magnetic field strength, B, and the mass of the ion, m. However, since the magnetic field strength is constant, cyclotron frequency depends only upon the inverse of mass-to-charge. It is also important to note that the cyclotron frequency of an ion is independent of its velocity and hence its kinetic energy. This feature allows FTMS to achieve ultra-high resolution. Equation 2, the definition of angular frequency, illustrates that an increase in an ion's velocity, v, will be accompanied by an increase in the radius, r, of the cyclotron orbit.

$$\omega_c = \frac{v}{r} \qquad (2)$$

Aside from cyclotron motion, ions exhibit two other motions in FTMS – trapping motion and magnetron motion.[47] Ions traveling along the magnetic field axis experience no force from the field. Therefore, to hold the ions in the cell, a potential is applied to two trapping plates that lie perpendicular to the magnetic field. Ions undergo harmonic oscillation along the magnetic field axis between the two trapping plates. This is known as trapping motion. While cyclotron motion and trapping motion are not coupled, the magnetic and electric fields in the ICR cell combine to produce magnetron motion. Magnetron motion is a precession of the center of the cyclotron motion around the center of the analyzer cell. Magnetron frequencies (1-100 Hz) are much lower than cyclotron frequencies (10 kHz-10 MHz).

FTICR instruments operate quite differently compared to most other mass analyzers. Most mass analyzers, including sector, quadrupole, and time-of-flight instruments, have their primary functions of ion formation, mass analysis, and ion detection occurring at the same time in different regions of the mass spectrometer. In FTMS, all of these events occur in the same space, the analyzer cell, at different times. A simple FTMS experiment consists of four separate events – ion quench, ion formation, ion excitation, and ion detection. The quench event simply clears the analyzer cell of any ions from a previous experiment. Ion formation (by MALDI for instance) may take place in the cell or just outside the cell with ions being transferred in prior to excitation.

Once ions are trapped in the cell, ion excitation occurs. Ions undergoing cyclotron motion are capable of absorbing energy from a radio frequency (rf) electric field applied to two plates of the analyzer cell, provided the frequency of the field matches the cyclotron frequency of the ions. The absorbed energy increases the velocity of the ion causing the ion to spiral outward thus increasing its cyclotron radius (Figure 1.5).[47] Ions of the same mass-to-charge are excited coherently and travel around the cell in a tightly grouped packet. Ions that are not excited remain at the center of the cell. However, broadband excitation using a rapid frequency sweep allows ions of many masses to be excited at the same time.

Unlike many other mass analyzers, ion detection in FTMS is nondestructive. Ions orbiting the cell in discrete mass-to-charge packets induce a current known as an image current (Figure 1.5). As a positive packet of ions approaches one detect plate, electrons are attracted to this plate causing a momentary current. As the ion packet continues around toward the opposite plate, the direction of the external electron flow is reversed resulting in the sinusoidal image signal. The magnitude of the image signal, or transient, depends upon the number of ions in the packet and the frequency of the mass-to-charge

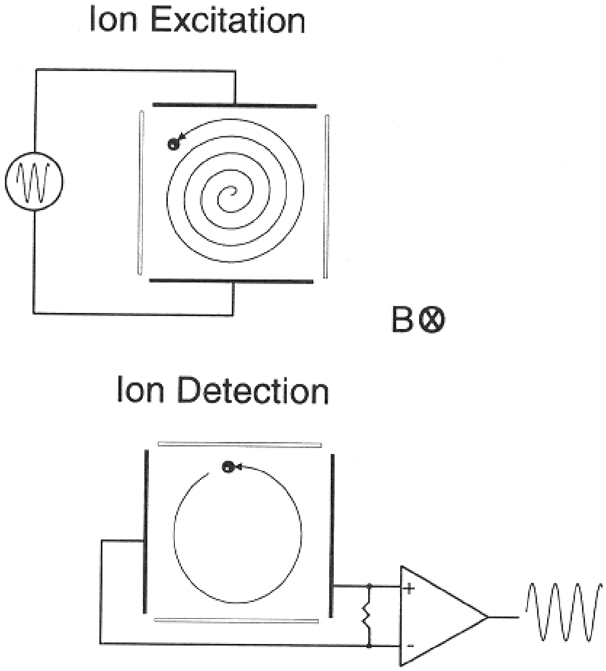

Figure 1.5 An illustration of ion excitation and detection in a Fourier transform mass spectrometry (FTMS) analyzer cell. The image was extracted from reference 47.

ratio of the ions in the packet. The frequency of the detected transient is equal to the difference between the cyclotron and magnetron frequencies of the ion.

The induced image current decays over a period of time (tenths of a second to several seconds) as the coherence of the ion packet deteriorates due to collisions with ions and neutrals in the analyzer cell. The image current is converted to a voltage that is acquired to provide the time-domain signal to be Fourier transformed to a frequency-domain signal. The frequency-domain spectrum can then be converted to a mass spectrum using a calibration formula derived from the cyclotron frequency equation shown below in Equation 3, where m is mass, f_{obs} is the observed ion frequency, and a and b are calibration constants.[48]

$$m = \frac{a}{f_{obs}} + \frac{b}{f_{obs}^2} \tag{3}$$

Equation 4 shows that mass resolution, R, increases proportional to the duration of the transient, T.[47]

$$R = \frac{\omega_c T}{2} \tag{4}$$

Figure 1.6 illustrates the importance of transient length upon mass resolution. The top mass spectrum corresponding to a 1.2 second transient is clearly more resolved than the mass spectrum arising from the 0.15 second transient. Therefore, FTMS experiments are performed at ultra-high vacuum to reduce the collision frequency.

PROTEOMICS

Proteomics is the systematic analysis of the proteins expressed by a cell or tissue type at a given time or under certain environmental conditions.[49] By examining the gene

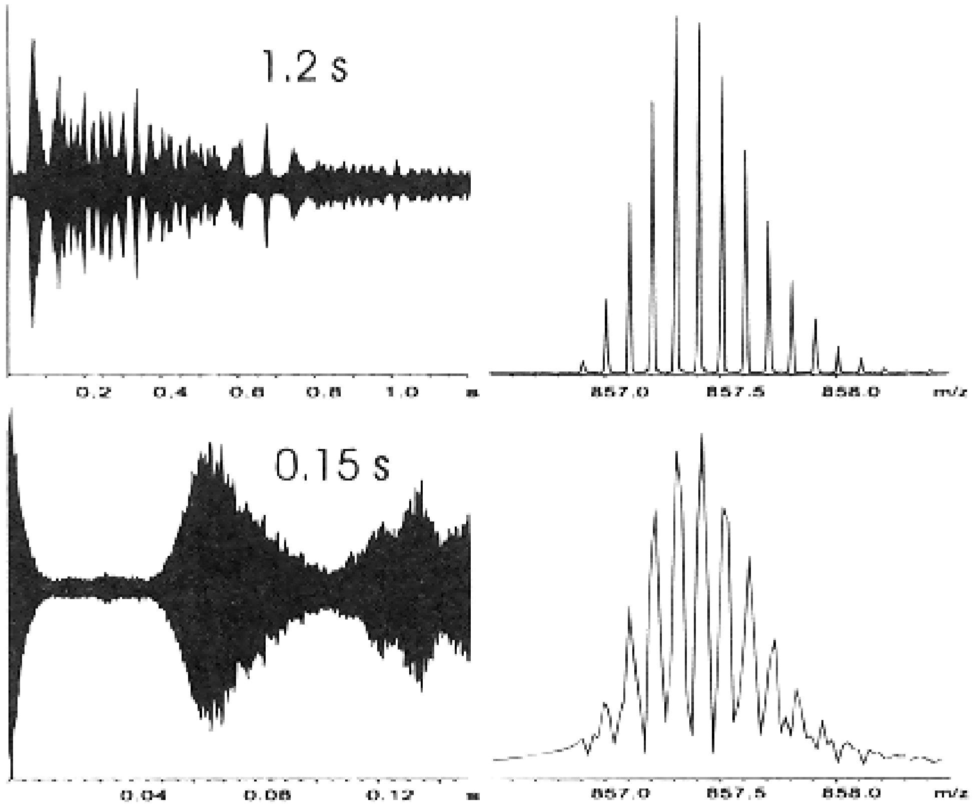

Figure 1.6 An illustration of the relationship between transient duration and mass resolution in Fourier transform mass spectrometry (FTMS). The image was extracted from reference 47.

products rather than the genes themselves, proteomics possesses the potential to unlock the information contained within genomic sequences including the structure, function, and regulation of biological processes.[50] While gene expression can be well characterized at the mRNA level due to the development of cDNA array technology, it is necessary to explore gene expression at the protein level.[51] Proteins catalyze essentially every biological function. Changes in protein concentration in response to stimuli are often achieved through proteolytic activation and post-translational modifications not closely related to gene expression and therefore cannot be detected by simply examining mRNA levels.[52] Assays of mRNA provide the rate at which a gene is transcribed, but do not necessarily provide insight into the identity or concentration of protein in the biological system since there is not always a good correlation between mRNA and protein levels.[53] In addition, many biological functions are the result of protein assemblies that cannot be examined at the gene level.[54]

The sudden boom in proteomics research is propelled by a number of factors - the heightened interest in genomics and genome sequencing, the advent of thorough, searchable databases that house sequence information, and the development of biological mass spectrometry for protein identification.[55] Near exponential growth of genome sequence information has arisen from advances in DNA sequence technology, robotic sample handling, and computer technology. Currently, the complete DNA sequence of over 20 organisms have been published and over 100 more are nearing completion.[56] To support the vast amount of genomic data being uncovered, advanced bioinformatic tools are being developed to organize, assimilate, and analyze data.

The analysis of proteins was revolutionized by the discovery of two MS ionization techniques in the 1980s, a discovery that earned the primary researchers the Nobel Prize. MALDI and ESI allow for large, nonvolatile proteins to be ionized without significant fragmentation. ESI made an immediate impact on protein chemistry since it can be coupled with liquid-phase separation techniques and because it produces multiply charged analyte ions. MALDI has also been used extensively in protein analysis since this method gives easy-to-interpret spectra from the singly charged analyte ions it produces, and it is relatively unaffected by interferences encountered in protein chemistry matrices (i.e. salts and detergents).[50]

Traditional Proteomics. For years, proteomic studies have relied on two-dimensional gel electrophoresis (2D-GE) as its sole method of analyzing the total protein complement of a cell or tissue. In 2D-GE, proteins are separated by two independent variables – isoelectric point (pI) and molecular weight – generating a "map" of hundreds to thousands of proteins.[57] The proteins are then visualized, excised from the gel, digested into peptides, and identified by Edman degradation or mass spectrometry (Figure 1.7). MS has become the preferred method of protein identification in speed, sensitivity, and reliability.[58] Using peptide mass fingerprinting, masses of peptides from proteins of interest are measured experimentally and compared with masses of peptides predicted by the amino acid sequence in protein databases.[59] Tandem mass spectrometry (MS/MS) can be used to yield additional information in the database search for a peptide. In MS/MS, a precursor peptide ion is selected and then fragmented to obtain at least partial amino acid sequence information.[60]

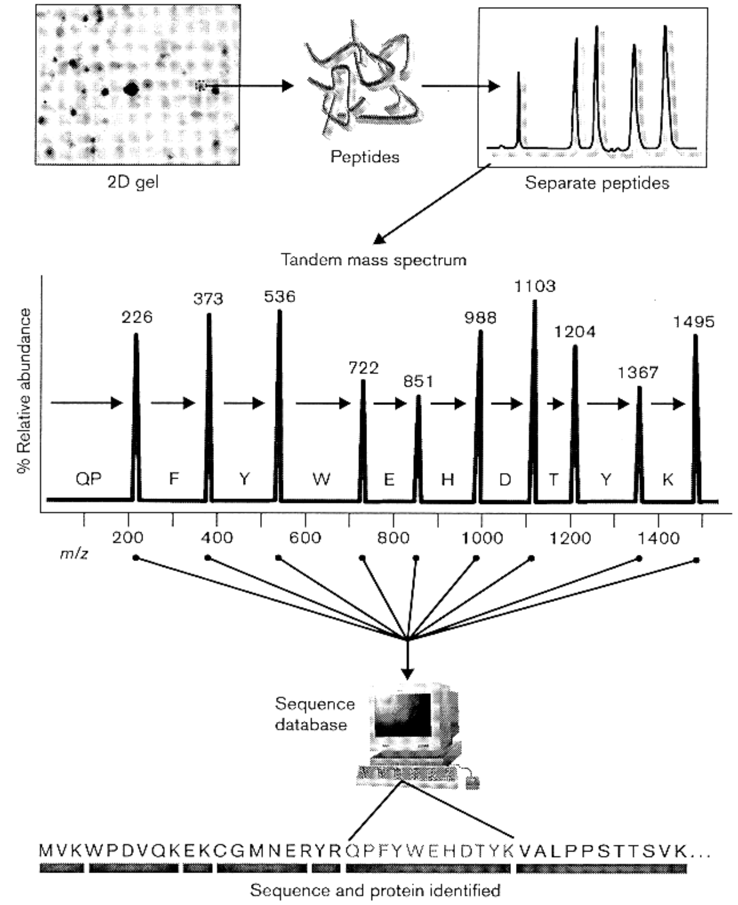

Figure 1.7 A schematic of protein identification by 2-dimensional gel electrophoresis (2D-GE) and tandem mass spectrometry (MS/MS). The image was adapted from reference 57.

While 2D-GE is useful for cataloguing proteins, it is not a global proteomic technique. Current 2D-GE visualization techniques detect only the most abundant proteins. An in-depth study of the yeast proteome showed that protein identification is biased towards high expression proteins, despite the fact that low expression proteins account for over half of the yeast proteome.[61] Therefore, low copy number proteins that have key roles in many biological pathways, such as regulatory proteins and transcription factors, cannot be identified. There are also difficulties associated in analyzing extremely acidic or basic proteins, very large or very small proteins, and hydrophobic proteins such as membrane proteins.[62,63] 2D-GE is also labor-intensive, time-consuming, and difficult to automate. These numerous limitations demonstrate the need for alternative proteomic techniques.

High-Throughput Proteomics. The majority of current proteomic research is focused upon the development of large-scale, rapid, and robust techniques for the analysis of proteins from complex biological mixtures. From this research, two different, but complementary methods of proteomic analysis by mass spectrometry have emerged – "top down" and "bottom up" proteomics. The "top down" method, first introduced by Fred McLafferty, involves the characterization of intact proteins using MS/MS and/or accurate mass measurements in order to generate sequence information.[64] The "top down" method is useful in the detection of post-translational modifications of proteins, but it is limited in the number of proteins that can be detected due to limited dynamic range and the scarcity of bioinformatic tools designed to analyze this type of data.[65]

The more versatile "bottom up" approach has been the focus of considerable interest due to its ability to rapidly identify large numbers of proteins and its ability to be easily adapted to quantitative methods. “Bottom up” proteomics, also known as shotgun proteomics, is based upon the enzymatic digestion of whole cellular fractions, typically followed by one or two-dimensional chromatography prior to MS analysis.[63] As opposed to 2D-GE where each protein is digested separately, only one batch digestion of the entire proteome is performed in shotgun proteomics. This method allows for the identification of low abundance proteins since peptides present at the sub-femtomole level can be handled as a mixture much easier than they can be excised from a gel and digested individually.[66] Therefore, shotgun proteomics offers a more comprehensive, simple, and rapid approach to processing the proteome of an organism.

The challenge of shotgun proteomics lies in the identification of proteins from their individual peptides. Batch digestion of even a simple proteome produces a complex mixture of thousands of peptides. In typical peptide mass fingerprinting experiments, like those performed to identify proteins in 2D-GE, each protein is digested individually. The protein can easily be identified since the majority, if not all, of the peptides present in the resulting mixture arise from a single protein. In order to identify a protein from a single peptide in a complex batch digest, more information needs to be known about an individual peptide than is required in a typical peptide mass fingerprinting experiment. There are two MS methods to elicit information from a single proteolytic fragment – tandem mass spectrometry or accurate mass measurement.

Tandem mass spectrometry has been the choice of many research groups in the identification of proteins from a single peptide. MS/MS yields partial amino acid sequence information of a peptide of interest. Therefore, proteins can be identified using the combination of the intact peptide mass and the masses of peptide fragments that arise from established peptide ion dissociation mechanisms. A research group headed by John Yates III has effectively utilized MS/MS to identify proteins from batch digests of whole cell lysates. Using their technique, multidimensional protein identification technology (MUDPIT), peptides are first separated by strong cation exchange (SCX) chromatography followed by RP-HPLC.[63] The peptides then elute directly into an ESI source coupled to a mass spectrometer capable of MS/MS. Using MUDPIT, the Yates group analyzed the whole-cell lysate of yeast and was able to identify 1484 proteins from the 5540 unique peptides generated.[67] While this MS/MS technique worked very well for shotgun proteomics, the use of accurate mass measurement for protein identification is thought to provide higher throughput.[56,68,69]

Accurate mass measurements (greater than 10 ppm mass accuracy) provide a different type of constraint for database searching than MS/MS.[70] In peptide mass fingerprinting, the more accurately masses are measured and the more overall peptide masses that can be linked to the same protein, the greater the confidence in the final identification of the protein.[71,72] However, an accurate mass measurement of a single peptide cannot alone provide enough information to positively identify a protein.[70] This point is illustrated by Figure 1.8, which shows the number of proteins from *Escherichia coli* that can give rise to specific tryptic peptides at a mass accuracy of 1, 10, and 100 ppm. The data for Figure 1.8 was generated by calculating the masses of all tryptic peptides for the entire *E. coli* proteome. Overall, as mass accuracy increases, the search specificity increases significantly. However, tryptic digestion produces a majority of peptides under 1000 mass units. Even at the high mass accuracy of 1 ppm that FTMS is capable of achieving, many peptide masses cannot conclusively identify a single protein.

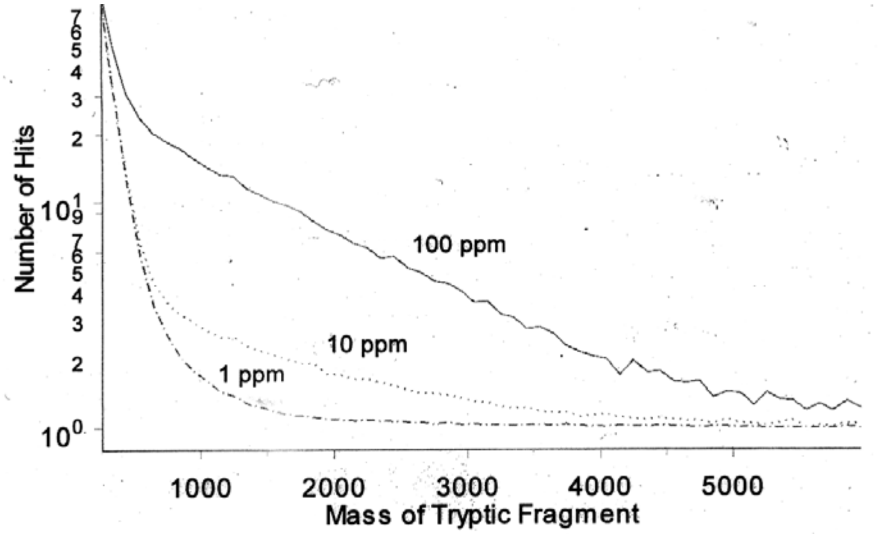

Figure 1.8 A diagram of search specificity for *Escherichia coli* proteins identified from its tryptic peptides measured at 1, 10, and 100 ppm mass accuracy.

The confidence in protein identification from a single accurate peptide mass can be increased if additional information is known about the peptide. As shown in Figure 1.9, search specificity can be increased significantly with knowledge that a certain low abundance amino acid, like cysteine, is present in the peptide of interest.[70] At even a modest mass accuracy of 10 ppm, the majority of the cysteine-containing peptides from *E.* coli can positively be assigned to a unique protein. For more complex organisms with larger proteomes, a higher degree of mass accuracy may be required.

Quantitative Proteomics. Most proteomic studies carried out to date have involved the simple identification of the proteins present in an organism. However, the protein complement of a cell is not static. In order to understand the dynamics of biological processes it is critical to obtain quantitative comparisons of the cell's protein composition in response to environmental stimuli. Therefore, proteomics is evolving to encompass quantitative studies to complement the already established qualitative studies. Quantitative proteomics has been accomplished with 2D-GE with limited success due to its complexity and limited dynamic range.[73] Quantitation is attempted by preparing two 2D-gels, a stressed and an unstressed system of the same organism.[74,75] A computer program then compares the intensity of the protein spots in order to determine if protein expression was up or down regulated in the organism.

Recently, alternatives to 2D-GE have been developed for quantitating the proteome of an organism. These methods rely on stable-isotope labeling – the addition of a chemically identical form of the analyte containing stable heavy isotopes (^{2}H, ^{13}C, ^{15}N, ^{18}O) to the sample to act as internal standards. Protein quantitation using stable-isotope labeling technology has been successfully reported by many research groups.[56,76-77] In

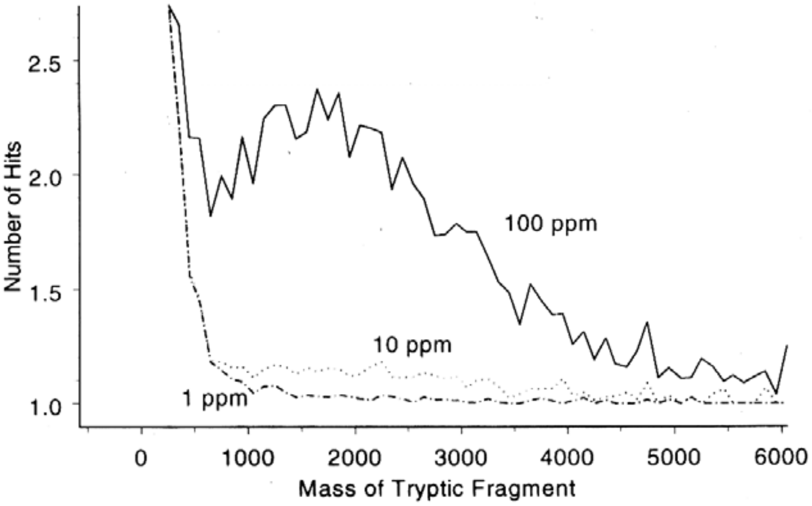

Figure 1.9 A diagram of search specificity for *Escherichia coli* proteins identified from its cysteine-containing tryptic peptides measured at 1, 10, and 100 ppm mass accuracy.

MS, the ionization efficiencies are highly variable for peptides so the only feasible internal standard is the exact same peptide labeled with a stable isotope.[78] Quantitation is possible by comparing an unstressed and a stressed cell type. One sample serves as the control while the other sample contains the same proteins but with different abundances due to the applied stress. Following digestion all peptides exist as pairs with a predictable mass difference due to the presence of the heavy isotope in the stressed sample. The relative abundance of the protein in the original cell lysate is inferred by measuring the intensity ratio between the peptide peak pairs.[79] Stable-isotope labeling of proteins can occur either endogenously using isotope-enriched media or exogenously using isotope-enriched reagents.

Quantitation using isotopically enriched nutrients during cell growth allows proteins to be encoded with stable isotopes at particular amino acid residues. Researchers under the direction of Richard Smith have successfully examined endogenous labeling using ^{15}N-enriched media.[56,76,80] Using this method, the organism of interest is grown on both ^{15}N-enriched media (>98%) and unenriched media. Following cell lysis, the two protein pools are combined, batch digested with trypsin, and analyzed by MS. The mass spacing between peptide peak pairs depends upon the number of nitrogen atoms in the peptide. Every amino acid residue in a peptide contains at least one nitrogen atom. Additionally, the side chains of arginine, asparagine, glutamine, histidine, lysine, and tryptophan contain one or more nitrogen atoms. Therefore, typical tryptic peptide peak pairs have spacing between 5-20 mass units. While not knowing the exact spacing may initially complicate mass spectral analysis, once the correlation is deduced, then the number of nitrogen atoms present in the peptide is known as well. This information assists in the identification of the parent protein. This very simple quantitative procedure is compatible with shotgun proteomic strategies for the comparative analysis of complex peptide mixtures.

Another novel stable isotope labeling scheme involves the incorporation of ^{18}O into tryptic peptides.[77] In an experiment performed by Catherine Feneslau and co-workers, one pool of proteins from adenovirus is digested by trypsin in ^{18}O-enriched water (95+% enrichment), while another pool of adenovirus proteins is digested under normal conditions. The proteins digested in ^{18}O-enriched water incorporate two ^{18}O atoms in their C-terminal carboxyl groups. Therefore, the peptides digested in ^{18}O-enriched water weigh 4 mass units more than the corresponding peptides digested in the unenriched water. The two pools are then mixed analyzed by MS to obtain quantitative data. This simple quantitative procedure is also compatible with quantitative shotgun proteomic strategies.

In 1999, Steven Gygi and co-workers developed a promising exogenous stable-isotope labeling approach using isotope coded affinity tags (ICATs).[81] The isotope coded affinity tags are incorporated into proteins of interest following extraction from the cell lysate. ICATs have three key functional parts – the biotin affinity tag, the linker, and the thiol-specific reactive group (Figure 1.10). For protein quantitation, ICAT reagents exist in heavy and light forms. In the heavy form the linker is eight-fold deuterated, while the light form is not isotopically labeled. Basic ICAT procedure calls for cells to be grown under two types of conditions and the proteins extracted (Figure 1.11). The proteins of interest are denatured and disulfide bonds are reduced to free sulfhydryls. Then one set of proteins is reacted with the light reagent while the other set is reacted with the heavy

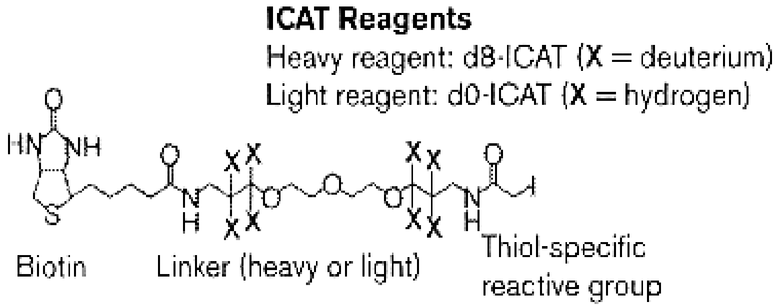

Figure 1.10 An illustration of isotope-coded affinity tag (ICAT) reagents for protein quantitation. The image was adapted from reference 81.

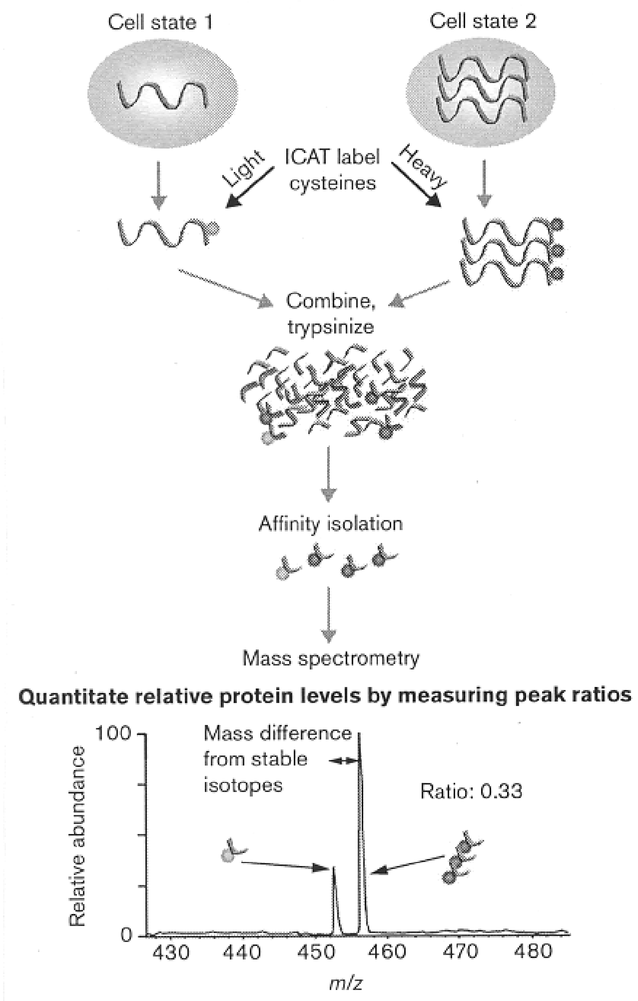

Figure 1.11 An overview of quantitative isotope-coded affinity tag (ICAT) analysis. The image was adapted from reference 81.

reagent. The thiol-specific reactive group binds to the free sulfhydryls of cysteines in the protein. The two samples are then mixed together, excess ICAT reagent is removed, and the proteins are digested into peptides by the protease trypsin. An avidin affinity separation is then performed in order to isolate the ICAT peptides. The biotin affinity tag present on the ICAT reagent binds to avidin thereby selectively isolating cysteine-containing peptides. After releasing the peptides from the avidin, quantitation and identification are performed by LC-MS/MS. Since each ICAT pair is nearly chemically identical, the pairs should essentially coelute from the LC and be found 8 mass units apart in a mass spectrum.

Quantitative proteomics using ICAT technology has many advantages. Unlike endogenous labeling using isotope-enriched media, ICAT analysis can examine any cell or tissue type grown under any possible condition. Also, since labeling occurs exogenously, protein levels are not altered by the presence of the isotope like they commonly are in endogenous labeling.[73] By mixing proteins following ICAT labeling, all proteins undergo the same digestion and separation procedures.[81] Therefore, any change in the abundance ratio of the proteins is due to the conditions under which the cells were grown. In addition, by selectively isolating cysteine-containing peptides, the complexity of the mixture to be analyzed by MS is reduced, and an additional database search constraint (the presence of a cysteine) is available identify the protein.[82] Finally, ICAT analysis is also able to quantitate proteins that cannot be analyzed by 2D-GE, including low mass and low abundance proteins.[78]

Many disadvantages are also encountered in ICAT analysis. ICAT only analyzes cysteine-containing peptides (roughly 15% of all peptides in a tryptic digest). Proteins that contain no cysteine residues (typically 10-15% of all expressed proteins) cannot be analyzed at all. The possibility does exist however to develop ICAT reagents to selectively isolate proteins by derivitization of other amino acids, such as histidine.[73,83] A bigger problem lies in the dependence of biotin-avidin binding to isolate cysteine-containing peptides. Difficulty is encountered in completely removing non-specifically, but tightly bound contaminants during the washing steps. Elution of all cysteine-containing peptides from the avidin column due to irreversible binding is also problematic.[84] A final limitation of ICAT is the significant chromatographic shift that results from the inequivalence of the light and heavy forms of the ICAT reagent. This complicates measurements of peptide abundances allowing quantitation of changes in protein expression with only 25% accuracy.[81]

The combination of exogenous stable-isotope labeling with endogenous stable isotope labeling can achieve even greater protein identification and quantitation specificity. Richard Smith and coworkers have analyzed both bacterial and mammalian proteomes using a combination of cysteine-affinity tags and ^{15}N-metabolic labeling.[85] In the experiments, cells are grown up on both ^{15}N-enriched media and unenriched media. The separate proteome pools are mixed and digested with trypsin. The complexity of the peptide mixture is then reduced by isolating cysteine-containing proteins using affinity tags similar to the “light” tags used in the ICAT procedure. This alternative, cost-effective high-throughput proteomic technique shows promising results without the significant isotope-dependent chromatographic shifts that plague quantitation by ICAT analysis.

Scope of This Thesis. Following a chapter describing the sample preparation and instrumentation used in performing the research, two experiments are detailed involving the development of shotgun proteomics using accurate mass measurements for protein identification. In Chapter 3, a LC-MALDI-FTMS based method for high-throughput proteomics is presented. In this study, the method is first tested using a standard protein digest and then is applied to a complex biological sample, the low molecular weight proteins of the archaeon *Methanococcus maripaludis*. An improvement upon shotgun proteomics using accurate mass measurements is described in Chapter 4. A novel technique called mass defect labeling is described and tested using a standard peptide. In the final chapter of this thesis, conclusions and future directions of proteomics research are addressed.

REFERENCES

1. Ehring, H., Karas, M., and Hillenkamp, F. *Org. Mass Spectrom.* **1992,** *27*, 472-480.
2. Wilson, C. M. In *Methods of Enzymology*; Hirs, C. H. W., and Timasheff, S.N., Ed., 1983; Vol. 91.
3. McLafferty, F. W. *Interpretation of Mass Spectra*; University Science Books: Sausalito, CA, 1993.
4. Munson, B., and Field, F.H. *J. Am. Chem. Soc.* **1966,** *88*, 2621.
5. Barber, M., Bordoli, R.S., Elliot, G.J., Sedgwick, R.D., and Tyler, A.N. *Anal. Chem.* **1982,** *54*, 645A.
6. Macfarlane, R. D., and Torgerson, D.F. *Science* **1976,** *191*, 930.
7. Beckey, H. D. *Principles of Field Ionization and Desorption Mass Spectrometry*; Pergamon, 1977.
8. Karas, M., and Hillenkamp, F. *Anal. Chem.* **1988,** *60*, 2299.
9. Tanaka, K., Waki, H., Ido, Y., Akita, S., Yoshida, Y., and Yoshida, T. *Rapid Commun. in Mass Spectrom.* **1988,** *2*, 151.
10. Suizdak, G. *Mass Spectrometry for Biotechnology*; Academic Press: San Diego, CA, 1996.
11. Beavis, R. C. *Org. Mass Spectrom.* **1992,** *27*, 653-659.
12. Knochenmuss, R., Dubois, F., Dale, M.J., and Zenobi, R. *Rapid Commun. in Mass Spectrom.* **1996,** *10*, 871-877.
13. Strupat, K., and Bridson. N.J. *J. Phys. D: Appl. Phys.* **1993,** *26*, 442-447.

14. Beavis, R. C., and Chait, B.T. *Rapid Commun. in Mass Spectrom.* **1989,** *3*, 432-435.

15. Beavis, R. C., Chaudhary, T., and Chait, B.T. *Org. Mass Spectrom.* **1992**, 156-158.

16. Overberg, A., Karas, M., and Hillenkamp, F. *Rapid Commun. in Mass Spectrom.* **1991,** *5*, 128.

17. Zhao, S., Somayajula, K.V., Sharkey, A.G., and Hercules, D.M. *Anal. Chem.* **1991**, 450.

18. Cornett, D. S., Duncan, M.A., and Amster, I.J. *Anal. Chem.* **1993,** *65*, 53.

19. Hillenkamp, F., Karas, M., Beavis, R.C., and Chait, B.T. *Anal. Chem.* **1991,** *63*, 1193A.

20. Tang, K., Allman, S.L., Jones, R.B., and Chen, C.H. *Org. Mass Spectrom.* **1992,** *27*, 1389.

21. Cornett, D. S., Duncan, M.A., and Amster, I.J. *Org. Mass Spectrom.* **1992,** *27*, 831.

22. Nelson, R. W., Thomas, R.M., and Williams, P. *Rapid Commun. in Mass Spectrom.* **1990,** *4*, 348.

23. Castoro, J. A., and Wilkins, C.L. *Anal. Chem.* **1993,** *65*, 2621.

24. Li, Y. Z., Hunter, R.L., and McIver, R.T. *International Journal of Mass Spectrometry and Ion Processes* **1996,** *158*, 175.

25. Zhang, W. Z., Niu, S.F., and Chait, B.T. *J. Am. Soc. Mass Spectrom.* **1998,** *9*, 879.

26. Niu, S. F., Zhang, W.Z., and Chait, B.T. *J. Am. Mass Spectrom.* **1998,** *9*, 1.

27. Budnick, B. A., Jensen, K.B., Jorgensen, T.D., Hasse, A., and Zubarev, R.A. *Rapid Commun. in Mass Spectrom.* **2000,** *14*, 578.

28. Laiko, V. V., Baldwin, M.A., and Burlingame, A.L. *Anal. Chem.* **2000,** *72*, 652.

29. O'Connor, P. B., and Costello, C.E. *Rapid Commun. in Mass Spectrom.* **2001,** *15*, 1862.

30. Salehpor, M., Perera, I., Kjellberg, J., Hedin, A., Islamian, M.A., Hakansson, P., and Sundqvist, B.U.R. *Rapid Commun. in Mass Spectrom.* **1994,** *3*, 2.

31. Dole, M., Mack, L.L., Hines, R.L., Mobley, R.C., Ferguson, L.D., and Alice, M.B. *J. Chem. Phys.* **1968,** *49*, 2240.

32. Yamashita, M., and Fenn, J.B. *J. Phys. Chem.* **1984,** *88*, 4451-4459.

33. Yamashita, M., and Fenn, J.B. *J. Phys. Chem.* **1984,** *88*, 4671-4675.

34. Taylor, G. *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences* **1964,** *280*, 383.

35. Taylor, G. *Proceedings of the Royal Society of London Series a-Mathematical and Physical Sciences* **1969,** *313*, 453.

36. Chen, R., Cheng, X., Mitchell, D.W., Hofstadler, S.A., Wu, Q., Rockwood, A.L., Sherman, M.G., and Smith, R.D. *Anal. Chem.* **1995,** *67*, 1159.

37. Henry, K. D., Quinn, J.P., and McLafferty, F.W. *J. Am. Chem. Soc.* **1991,** *113*, 7544.

38. Beu, S. C., Senko, M.W., Quinn, J.P., Wampler, F.M. III, and McLafferty, F.W. *J. Am. Mass Spectrom.* **1993**, 557.

39. Bardoli, R. S., Howes, K., Vickers, R.G., Bateman, R.H., and Harvey, D.J. *Mass Spectrom. Biomolec. Sci.* **1996,** *475*, 51.

40. Cottrell, J. S. *Org. Mass Spectro.* **1992,** *27*, 343.

41. Qin, J., Steenvoorden, R.J.,and Chait, B.T. *Anal. Chem.* **1996,** *69*, 4002.

42. Fannin, S. T., Wu, J., Molinski, T., and Lebrilla, C.B. *Anal. Chem.* **1995,** *67*, 3788.

43. Carroll, J. A., Penn, S.G., Fannin, S.T., Wu, J., Cancilla, M.T., Green, M.K., and Lebrilla, C.B. *Anal. Chem.* **1996,** *68*, 1798.

44. Sommer, H., Hipple, H.A., and Thomas, J.A. *Phys. Rev.* **1951,** *82*, 697.

45. Comisarow, M. B., and Marshall, A.G. *Chem. Phys. Lett.* **1974,** *25*, 282-283.

46. Comisarow, M. B., and Marshall, A.G. *Chem. Phys. Lett.* **1974,** *26*, 489-490.

47. Amster, I. J. *J. Mass Spec.* **1996,** *31*, 1325-1337.

48. Gross, M. L., and Rempel, D.L. *Science* **1984,** *226*, 261.

49. Nyman, T. A. *Biomol. Eng.* **2001,** *18*, 221-227.

50. Aebersold, R., and Goodlett, D. R. *Chem. Rev.* **2001,** *101*, 269-295.

51. Lashkari, D. A., DeRisi, J.L., McCusker, J.H., Namath, A.F., Gentile, C., Hwang, S.Y., Brown, P.O., and Davis, R.W. *Proc. Natl. Acad. Sci. USA* **1997,** *94*, 1314-1334.

52. Celis, J. E., Kruhoffer, M., Gromova, I., Frederiksen, C., Ostergaard, M., Thykjaer, T., Gromov, P., Yu, J., Palsdottir, H., Magnusson, N., and Orntoft, T.F. *FEBS Lett.* **2000,** *480*, 2-16.

53. Gygi, S. P., Rochon, Y., Franza, B.R., and Aebersold, R. *Mol. Cell. Biol.* **1999,** *19*, 1720-1730.

54. Peters, E. C., Brock, A., Horn, D.M., Phung, Q.T., Ericson, C., Salomon, A.R., Ficarro, S.B., and Brill, L.M. *LC GC Eur.* **2002,** *15*, 423-428.

55. Fields, S. *Science* **2001,** *291*, 1221-1224.

56. Conrads, T. P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L., and Smith, R.D. *Anal. Chem.* **2000,** *72*, 3349-3354.

57. Gygi, S. P., and Aebersold, R. *Curr. Op. Biotech.* **2000,** *4*, 489.

58. Andersen, J. S., and Mann, M. *FEBS Lett.* **2000,** *480*, 25-31.

59. Mortz, E., Vorm, O., Mann, M., and Roepstorff, P. *Biol. Mass Spectrom.* **1994,** *23*, 249.

60. Gerraet, K., and Vandekerckhove, J. *Electrophoresis* **2000,** *20*, 1145-1154.

61. Shevchenko, A., Jensen, O.N., Podtelejnikov, A.V., Sagliocco, F., Wilm, M., Vorm, O., Mortensen, P., Boucherie, H., and Mann, M *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 14440.

62. Smith, R. D. *Nat. Biotech.* **2000,** *18*, 1041-1042.

63. Wolters, D. A., Washburn, M.P., and Yates, J.R. *Anal. Chem.* **2001,** *73*, 5683.

64. Mortz, E., O'Connore, P.B., Roepstroff, P.Kelleher, N.L., Wood, T.D., McLafferty, F.W., and Mann, M. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 8264-8267.

65. Meng, F. Y., Cargile, B.J., Miller, L.M., Forbes, A.J., Johnson, J.R., and Kelleher, N.L. *Nat. Biotech.* **2001,** *19*, 952-957.

66. Martin, S. E., Shabanowitz, J., Hunt, D.F., and Marto, J.A. *Anal. Chem.* **2000,** *72*, 4266.

67. Washburn, M. P., Wolters, D., and Yates, J.R. *Nat. Biotech.* **2001,** *19*, 242.

68. Lipton, M. S., Pasa-Tolic, L., Anderson, D.J., Auberry, D.L., Battista, K.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie,

L.M., Moore, R.J., Romine, M.F., Shen, Y.F., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, L.K., Zhao, R., and Smith, R.D. *Proc. Natl. Acad. Sci. USA* **2002,** *99*, 11049-11054.

69. Smith, R. D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., Shen, Y.F., Conrads, T.P., Veenstra, T.D., and Udseth, H.R. *Proteomics* **2002,** *2*, 513-523.

70. Goodlett, D. R., Bruce, J.E., Anderson, G.A., Rist, B. Pasa-Tolic, L., Fiehn, O., Smith, R.D., and Aebersold, R. *Anal. Chem.* **2000,** *72*, 1112.

71. Fenyo, D., Qin, J., and Chait, B.T. *Electrophoresis* **1998,** *19*, 998-1005.

72. Zubarev, R. A., Hakansson, P., and Sundqvist, B. *Anal. Chem.* **1996**, 4060-4063.

73. Gygi, S. P., and Aebersold, R. In *Proteomics: A Trends Guide*; Elsevier Science Ltd.: London, 2000, pp 31-36.

74. Hochstrasser, D. F. *Clin. Chem. Lab. Med.* **1998,** *36*, 825.

75. Jungblot, P., and Wittmannliebold, B. *J. Biotechnol.* **1995,** *41*, 111.

76. Oda, Y., Huang, K., Cross, F.R., Cowburn, D., and Chait, B.T. *Proc. Natl. Acad. Sci. USA* **1999,** *96*, 6591-6596.

77. Yao, X. D., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. *Anal. Chem.* **2001,** *73*, 2836.

78. Gygi, S. P., Rist, B., and Aebersold, R. *Curr. Op. Biotech.* **2000,** *11*, 396.

79. Peng, J., and Gygi, S.P. *J. Mass Spec.* **2001,** *36*, 1083-1091.

80. Veenstra, T. D., Martinovic, S., Anderson, G.A., Pasa-Tolic, L., and Smith, R.D. *J. Am. Soc. Mass Spectrom.* **2000,** *11*, 78-82.

81. Gygi, S. P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. *Nat. Biotech.* **1999,** *17*, 994.

82. Mann, M. *Nat. Biotech.* **1999,** *17*, 954-955.

83. Ji, J., Chakraborty, A., Geng, M., Zhang, X., Amini, A., Bina, M., and Regnier, F. *J. Chrom.* **2000,** *745*, 197-210.

84. Moseley, M. A. *Trends Biotechnol.* **2001,** *19*, S6-S10.

85. Conrads, T. P., Alving, K., Veenstra, T.D. Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Pasa-Tolic, L., Udseth, H.R., Chrisler, W.B., Thrall, B.D., and Smith, R.D. *Anal. Chem.* **2001,** *73*, 2132.

CHAPTER 2

EXPERIMENTAL

The purpose of this chapter is to provide additional information concerning the instrumentation and sample preparation used in performing the high-throughput proteomics studies described in Chapters 3 and 4. The research presented in the following chapters share common techniques that are briefly outlined in their respective experimental sections. This chapter, however, offers a detailed examination of select analysis and sample preparation techniques.

INSTRUMENTATION

LC System. An Ultra Plus II MD capillary HPLC system (Micro-Tech Scientific Inc., Vista, CA) capable of performing both one-dimensional and two-dimensional separations is used to reduce the complexity of peptide mixtures prior to MS analysis (Figure 2.1). The LC system contains four pumps and a small volume static mixer allowing for two separate capillary scale chromatography gradients. However, in the research chapters that follow, only simple one-dimensional separations are utilized based on the complexity of samples analyzed.

Peptide mixtures are separated using a reverse-phase capillary HPLC column (150 μm i.d. x 15 cm) packed with C18 stationary phase (Figure 2.2). A peptide digest is injected into the 5 μL sample loop using a 50 μL syringe. A flow rate of 0.5 μL/min is used to elute components. Over the course of a 75-minute separation, the solvent gradient changes from 100 % water with 1 % acetic acid to 100 % acetonitrile. The final 5 minutes of the LC run is held at 100 % acetonitrile to ensure all peptides have eluted

Figure 2.1 A photograph of a Micro-Tech Ultra Plus II MD capillary HPLC system. The image was extracted from http://www.microlc.com/newweb/2Dsetup12.htm.

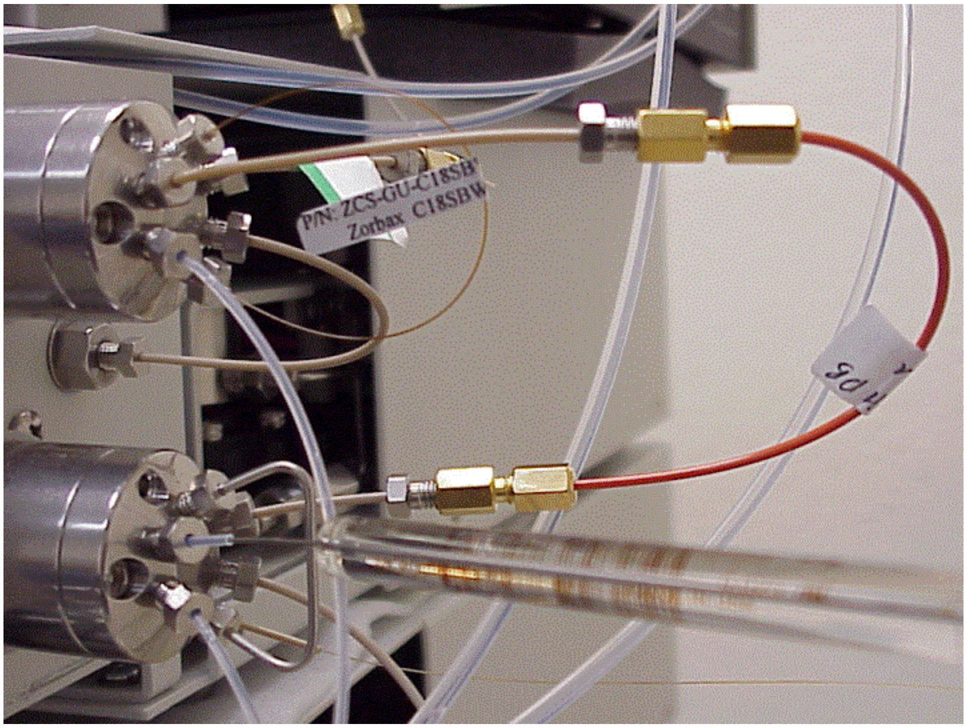

Figure 2.2 A photograph of a reverse-phase capillary HPLC setup. The image was extracted from http://www.microlc.com/newweb/2Dsetup13.htm.

from the column. The LC column eluent then flows through a capillary UV detector (250 nL volume) set at 214 nm. However, the UV detector can also be removed in order to decrease dead volume. The eluent from the capillary HPLC can be analyzed on-line with direct flow into an ESI source or off-line using an autosampler to collect fractions for MALDI analysis.

Autosampler. A Gilson 215 single-channel autosampler (Middleton, WI) is used to collect LC fractions and directly apply MALDI matrix to a multi-spot MALDI target (Figure 2.3). A 1-meter piece of 25 μm i.d. fused silica capillary is connected to the outlet of the reverse-phase capillary HPLC system. The other end of the fused silica capillary is threaded through the metal delivery tube of the autosampler. The fused silica capillary protrudes approximately 1 mm from the delivery tube. Deposition of an LC fraction occurs without the fused silica capillary ever touching the MALDI target surface. When a calculated volume is accumulated at the tip of the capillary, the tip is lowered and the droplet is transferred to a spot on the MALDI target. The capillary tip is then advanced to the next spot of the MALDI target. The low flow rate of capillary HPLC allows for the collection of typical MALDI sample volumes. The autosampler is programmed to collect 1 μL fractions in 2 minutes based on a 0.5 μL/min capillary LC flow rate. A single MALDI target has 49 spots, and the autosampler is equipped with racks to hold up to 20 MALDI targets. A typical one-dimensional LC run can be deposited on one MALDI target, while additional MALDI targets are necessary for a two-dimensional separation.

Following the completion of an HPLC run, the fused silica capillary is removed from metal delivery tube in order to deposit MALDI matrix. The autosampler is

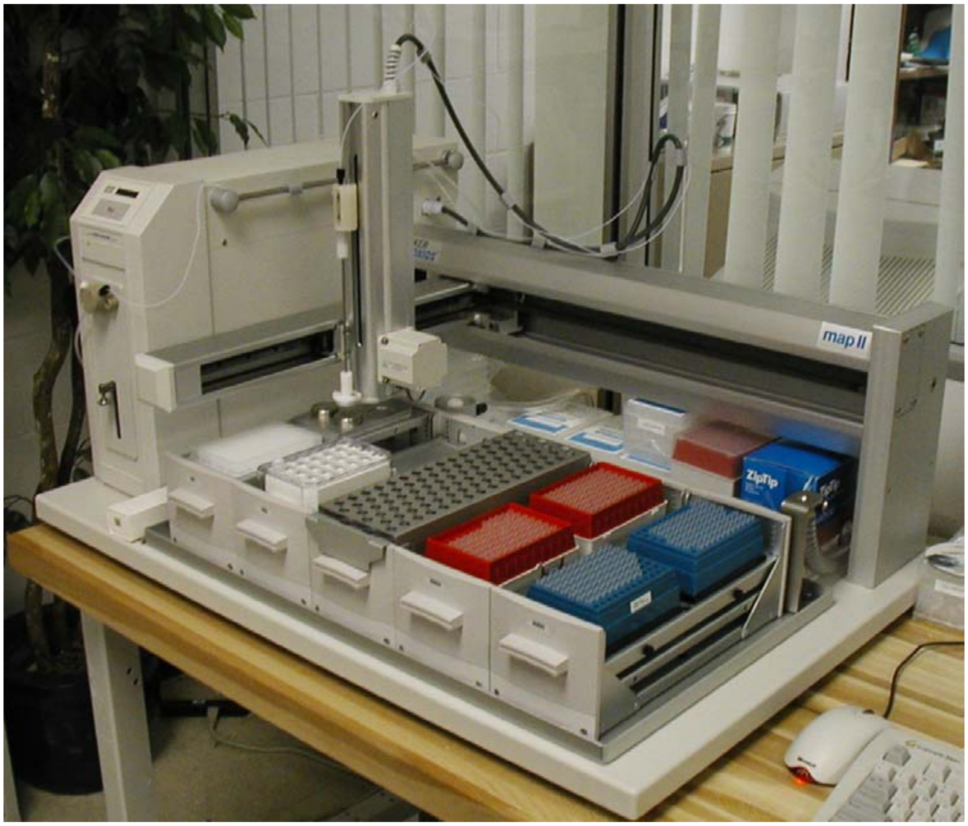

Figure 2.3 A photograph of a Gilson 215 single-channel autosampler.

programmed to pick up a disposable plastic tip, collect 1 μL of MALDI matrix from a vial, and deposit the matrix solution on a LC fraction. The LC fraction and matrix are mixed within the disposable plastic tip through aspiration and ultimately co-deposited on the MALDI target (Figure 2.4). The plastic tip is discarded, and the process is continued until MALDI matrix is applied to all LC fractions. Once the MALDI spots dry, FTMS analysis can be performed. The entire process involves limited human intervention ensuring reproducible sample deposition.

Intermediate Pressure MALDI-FTMS. High-throughput proteomic experiments were performed using a FTMS-130 intermediate pressure MALDI source at Bruker Daltonics, Inc. (Billerica, MA). A schematic of the MALDI source is shown in Figure 2.5. Once inside the source vacuum chamber, an X-Y manipulator remotely controls the 49-spot MALDI target. The manipulator positions the sample spot of interest within 0.2 mm accuracy of a 337 nm UV desorption laser. In a typical intermediate pressure MALDI experiment, a pulse valve is opened letting in argon gas to raise the pressure in the MALDI source to 10^{-3} torr. The MALDI target voltage is set at 22 V for the duration of the experiment. The laser is fired multiple times at the MALDI target (1-10 depending upon signal strength), and the pulse valve closes 100 msec after the final laser firing. A 100 msec delay for pump-down of the source occurs prior to ion extraction. Desorbed ions accumulate and pass through a hexapole. During the MALDI event, the hexapole exit lens is set to 23 V allowing ions to accumulate in the hexapole. However, throughout extraction the exit voltage is dropped to –8 V, forcing ions out of the hexapole. Ions exit the hexapole and are directed into the FTICR analyzer cell using

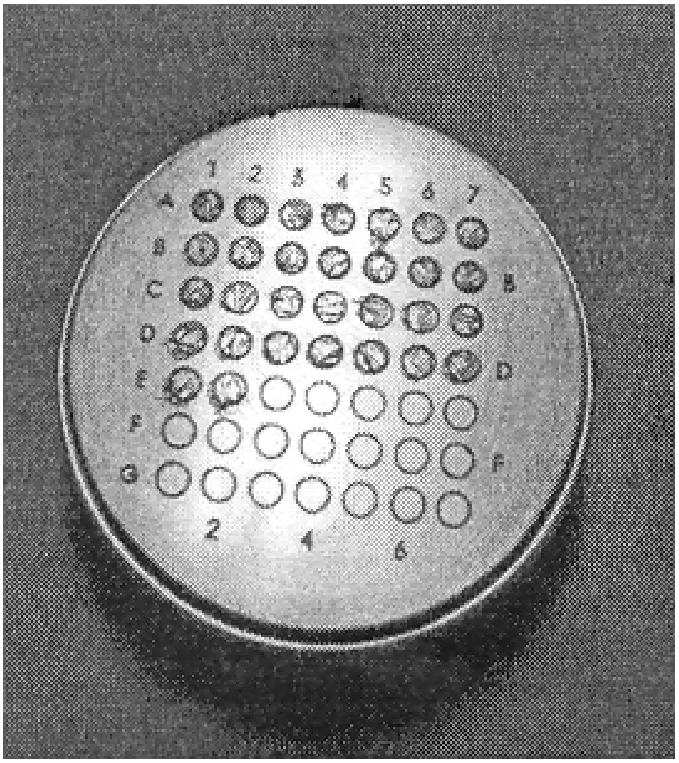

Figure 2.4 A photograph showing co-deposition of LC fractions and matrix on a 49-Spot MALDI target.

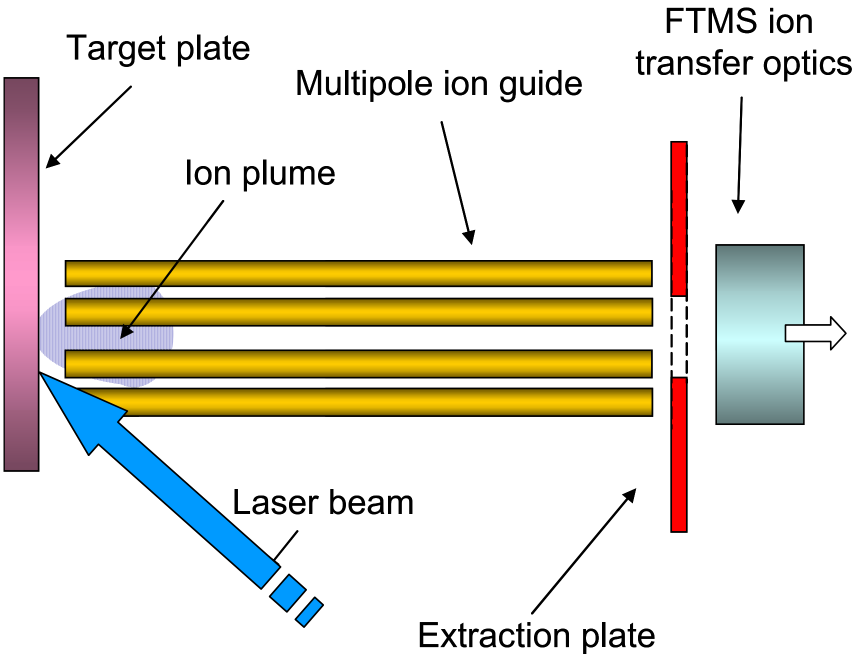

Figure 2.5 A schematic of a Bruker FTMS-130 intermediate pressure MALDI source.

electrostatic ion transfer optics where they are trapped using sidekick. Sidekick applies a voltage to the ions as they enter the FTICR cell through a small hole. The applied voltage takes the ions off-axis and expands their cyclotron radius causing them to be trapped in the FTICR cell for analysis. The IP MALDI source is coupled to a 7.0 Telsa Bruker BioApex Fourier transform ion cyclotron resonance mass spectrometer (Figure 2.6).

In-House FTICR Mass Spectrometers. High vacuum MALDI experiments were performed using a 4.7 Tesla Fourier transform ion cyclotron resonance mass spectrometer designed and fabricated at the University of Georgia (Figure 2.7).[1-6] The instrument is equipped with an internal MALDI source which forms ions adjacent to the ICR cell allowing for high trapping efficiency and increased sensitivity.[7] Samples are deposited on a 2.5 cm diameter titanium MALDI target. Once in the ultra-high vacuum chamber (10^{-9} to 10^{-10} torr), the target is positioned ~1 cm from the front trap of an open cylindrical FTICR cell using a magnetically coupled linear-rotary manipulator. A 355 nm Nd:YAG laser is focused through the cell and onto the MALDI target. Following the MALDI event, gated trapping captures the ions.[2,3,8] As the ions enter the cell, the potential on the back trap is raised to prevent the ions from passing through the cell. The back trapping potential reverses the direction of the ions, so the front trapping potential is raised equal to that of the back trap thereby holding the ions in the FTICR cell. A commercial data station controls the mass spectrometer (Ionspec Corp., Irvine, CA).

Additional proteomics experiments were performed using a 7.0 Tesla Bruker BioApex Fourier transform mass spectrometer with an external ESI source. The eluent of the capillary LC system flowing at 0.5 μL/min is coupled directly to the ESI source. A

Figure 2.6 A photograph of a FTMS-130 intermediate pressure MALDI source coupled to a 7.0 tesla Bruker BioApex Fourier transform mass spectrometer.

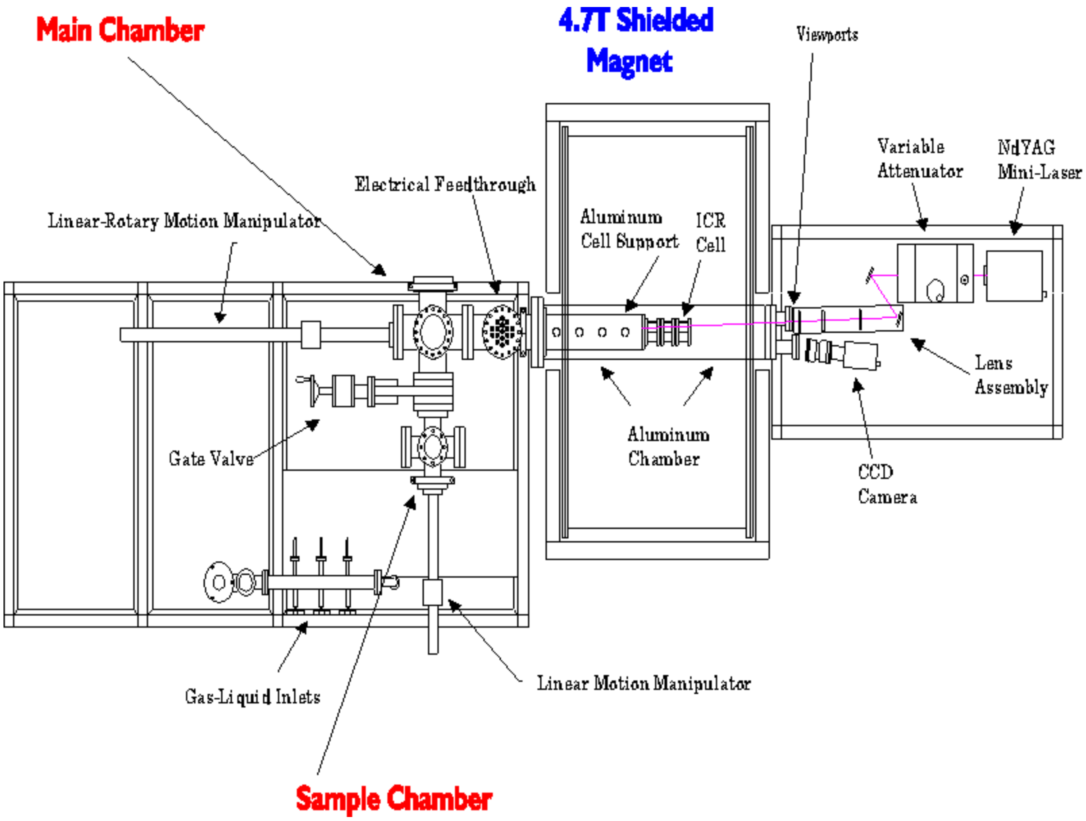

Figure 2.7 A schematic of a 4.7 tesla MALDI Fourier transform mass spectrometer.

~1.5 kV applied potential produces ions at atmospheric pressure. Ions pass through a skimmer as they enter the high vacuum chamber (10^{-9} to 10^{-10} torr) of the mass spectrometer. Ion accumulation occurs in a hexapole prior to entry into the FTICR cell. Ions are directed into a cylindrical infinity FTICR cell using electrostatic ion transfer optics where they are trapped using sidekick.

SAMPLE PREPARATION

Protein Digestion. Shotgun proteomics requires the digestion of proteins into peptides prior to mass spectral analysis. Standard proteins or mixtures of standard proteins are initially analyzed when testing a new shotgun proteomics procedure. Bovine serum albumin (BSA), ovalbumin, carbonic anhydrase, and myoglobin are just a few of the many commercially available proteins that make ideal analysis standards. The type of experiment often dictates the choice of standard protein. For example, in a cysteine labeling experiment such as ICAT analysis, it is necessary to choose a protein that contains cysteine, such as BSA. In our laboratory, standard protein mixtures are made at a concentration of 1 mg/mL in 100 mM ammonium bicarbonate buffer (pH 8). It is very important to choose a buffer that keeps proteins stable in solution and that is compatible with MS analysis.

Once the validity of a proteomic method is confirmed using a simple standard protein digest, a more complicated mixture such as a whole-cell lysate may be examined. Unlike standards, proteins obtained directly from biological sources often contain contaminants as a result of extraction and/or purification. Detergents, such as sodium dodecyl sulfate (SDS), or buffering agents, such as tris hydroxymethylaminoethane, are often present to maintain protein activity, prevent protein precipitation, and reduce adsorption to synthetic surfaces. Contaminants may interfere with protein digestion and/or MS analysis, so buffer exchange is typically performed prior to digestion to reduce the overall concentration of contaminants. In our laboratory, buffer exchange is accomplished by membrane filtration using 3 kDa molecular weight cut-off filters.

Once proteins are in a compatible buffer solution, protein denaturation and reduction of disulfide linkages is performed. Proteins are tightly bound globular structures. In order to effectively digest proteins into their constituent peptides, the globular structure must be unfolded. Denaturation is most commonly accomplished by simply heating the protein solution at 95 °C for 5-10 minutes. The addition of chaotropes, such as 6 M guanidinium or 8 M urea, also assists in the denaturation of proteins. Proteins that contain multiple cysteine residues are more tightly bound due to the formation of disulfide linkages. Reduction of the disulfide linkages to free sulfhydryls allows for further denaturation to occur. In addition, free sulfhydryls are required for any cysteine labeling experiments. In our laboratory, tris (2-carboxyethyl) phosphine (TCEP) is used in the reduction process rather than tributylphosphine (TBP) or dithiotheritol (DTT). TCEP has all the advantages of TBP and DTT (stoicheometric, quantitative, specific, fast) but is odorless (unlike DTT and TBP) and soluble in water (unlike TBP).

At this point in the digestion procedure, labeling of select amino acid residues may occur depending upon the proteomics experiment. Labeling reagents are added in excess to ensure complete reaction with the substrate. The presence of both excess labeling reagents as well as denaturation and/or reduction reagents can have an adverse effect upon protein digestion and/or mass spectral analysis. In our laboratory, sample purification is performed prior to protein digestion using gel permeation chromatography (GPC). In GPC, a protein solution is passed through a column packed with an inert, microporous stationary phase. Small molecules, such as reduction reagents, penetrate more pores and are retained in the stationary phase. Large molecules, such as proteins, pass swiftly through the column. Sephadex gel columns are manufactured in-house and used in conjunction with a swinging bucket centrifuge for GPC.

Finally, protein digestion is performed. Many chemicals and enzymes are commercially available for specific and non-specific protein cleavage. The most commonly used enzyme for specific protein cleavage is trypsin. With a few exceptions, trypsin cleaves on the carboxyl side of arginine and lysine residues. No cleavage occurs when the amino acid proline is at carboxylic side of lysine or arginine, and reduced cleavage occurs when acidic residues are present on either side of susceptible bond. In our laboratory, trypsin modified by reductive alklyation is used in order to prevent autodigestion. Typical protocol calls for a 20:1 to 100:1 (w/w) ratio of substrate-to-enzyme with digestion occurring overnight at 37 °C. Freezing the peptide solution ceases digestion.

MALDI Target. An important aspect of MALDI is the choice of an appropriate matrix. With many matrices available for analysis, factors to consider when choosing a matrix include analyte type, laser wavelength, and ease of sample preparation. 2,5-dihydroxybenzoic acid (DHB) was chosen as the matrix for all MALDI experiments. DHB is often used for peptide analysis. It is a relatively “cool” matrix allowing for little fragmentation of analyte molecules. DHB is also water-soluble and compatible with many solvents allowing for easy sample preparation. This sometimes creates problems, since it is possible to have too much matrix in a sample spot. As a result, overwhelming matrix signal and/or suppression of analyte ion formation can occur.[9,10] Typically, analyte is co-crystallized in a matrix solution in a 1000-to-1 matrix-to analyte ratio.[11]

DHB was recrystallized prior to MS analysis to increase the purity of the matrix. DHB was supersaturated with acetone and thoroughly mixed. The DHB solution was centrifuged and the supernatant was discarded. Formic acid was then added to the remaining solid, and the DHB solution was thoroughly mixed and refrigerated for 30 minutes. The mixture was centrifuged, and the supernatant was discarded. The recrystallized DHB was dried under vacuum.

1 M DHB was prepared fresh for every experiment in 50 % aqueous acetonitrile with 0.1 % v/v trifluoroacetic acid (TFA). Small amounts of TFA have been shown to increase protein solubility and improve crystallization.[12,13] The dried droplet technique was used to spot MALDI targets. 1 μL of a digested peptide solution is first spotted on the MALDI target, and 1 μL of 1 M DHB is then deposited on top of the analyte molecules. The two solutions are repeatedly aspirated in the tip of a disposable plastic pipet tip to thoroughly mix the solutions. The peptide solution and the matrix are then ultimately co-deposited on the MALDI target. The spots require ~15 minutes to dry prior to MALDI-FTMS analysis. This analyte and matrix deposition process occurs manually for high vacuum MALDI experiments, while the autosampler is used for intermediate pressure MALDI experiments.

Salts or other water-soluble contaminants that interfere with ion production are often found in proteolytic peptide digests. While MALDI is relatively tolerable of interferents, removal of these contaminants often aids mass spectral analysis. One sample clean-up method involves the immobilization of peptides on a nitrocellulose layer followed by rinsing with cold, deionized water.[14-16] In this procedure, 5 mg of nitrocellulose is dissolved in 0.5 mL of acetone, and 1 μL of the solution is applied to the MALDI target. Once dry, 1 μL of the contaminated peptide solution is deposited on top of the dried nitrocellulose spot. Peptides are noncovalently immobilized on the nitrocellulose film. Once dry, 1 μL of cold, deionized water is applied to the sample allowing for dissolution of the water-soluble contaminants. After 30 seconds, excess water is removed by aspiration. Finally, 1 μL of 1M DHB is spotted prior to MALDI-FTMS analysis. Besides removal of contaminants, this procedure offers the additional advantage of adding nitrocellulose as a co-matrix to provide collisional cooling during the initial laser desorption.

REFERENCES

1. Easterling, M. L., Pitsenberger, C.C., Kulkarni, S.S., Taylor, P.K., and Amster, I.J. *Int. J. Mass Spectrom. Ion Process.* 1996, *158*, 97.
2. Easterling, M. L., Pitsenberger, C.C., and Amster, I.J. *J. Am. Soc. Mass Spectrom.* 1997, *8*, 195.
3. Easterling, M. L., Mize, T.H., and Amster, I.J. *Int. J. Mass Spectrom.* 1997, *169*, 387.
4. Pitsenberger, C. C., Easterling, M.L., and Amster, I.J. *Anal. Chem.* 1996, *68*, 3732.
5. Pitsenberger, C. C., Easterling, M.L., and Amster, I.J. *Anal. Chem.* 1996, *68*, 4409.
6. Mize, T. H., and Amster, I.J. *Anal. Chem.* 2000, *72*, 5586.
7. Solouki, T., Marto, J.A., White, F.M., Guan, S.H., and Marshall, A.G. *Anal. Chem.* 1995, *67*, 4139.
8. Vartanian, V. H., and Laude, D.A. *J. Am. Soc. Mass Spectrom.* 1995, *6*, 812.
9. Tang, X. D., Driefuss, P.A., and Vertes, A. *Rapid Commun. in Mass Spectrom.* 1995, *9*, 1141.
10. Blais, J. C., Tessier, M., Bolbach, G., Remaud, B., Rozes, L., Guittard, J., Brunot, A., Merechal, E., and Tabert, J.C. *Int. J. Mass Spectrom. Ion Process.* 1995, *144*, 131.
11. Schurenberg, M., Schulz, T., Dreisewered, K., and Hillenkamp, F. *Rapid Commun. in Mass Spectrom.* 1996, *10*, 1873.

12. Hillenkamp, F., Karas, M., Beavis, R.C., and Chait, B.T. *Anal. Chem.* 1991, *63*, 1193A.
13. Beavis, R. C., and Chait, B.T. *Chem. Phys. Lett.* 1991, *181*, 479.
14. Preston, L. M., Murray, K.K., and Russell, D.H. *Bio. Mass Spectrom.* 1993, *22*, 544.
15. Zaluzec, E. J., Gage, D.A., Allison, J., and Watson, J.T. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 230.
16. Worrall, T. A., Cotter, R.J., and Woods, A.S. *Anal. Chem.* 1998, *70*, 750.

CHAPTER 3

A LC-MALDI-FTMS BASED METHOD FOR HIGH-THROUGHPUT PROTEOMICS

INTRODUCTION

In the burgeoning field of high-throughput proteomics, the development of viable alternatives to two-dimensional gel electrophoresis (2D-GE) has been the focus of many research groups.[1-10] The majority of current research involves the rapid analysis of an entire proteome via batch enzymatic digestion of whole-cell lysates or cellular sub-fractions.[1-9] This method, known as shotgun proteomics, produces a complex mixture of peptides. Proteins are then identified from their constituent peptides. Nearly all of the new methodologies accomplish this process using HPLC coupled on-line with ESI-MS and/or ESI-MS/MS.[1-8] Presented here is a LC-MALDI based method for high-throughput proteomics to complement the ESI methods already developed.

LC-MALDI. Batch digestion of an entire cellular fraction produces a very complex mixture of peptides. Even simple organisms, such as bacteria or archaea, express more than 1000 proteins at a given time. Tryptic digestion of the entire proteome yields several thousand to tens of thousands of peptides. A mass spectrum of the resulting mixture would be impossible to analyze. Peptide peaks would appear at every mass unit with significant overlap of both monoisotopic peaks as well as their corresponding isotope peaks. The ultra-high resolution capabilities of FTMS cannot be solely relied upon to differentiate all of the monoisotopic peptide peaks in a mixture of such complexity. Also, the FTICR analyzer cell has a finite charge capacity, and the dynamic range for any individual component will be limited when too many components are present in the cell at one time. Therefore, peptide mixtures must be simplified prior to mass analysis by using separation techniques. Reverse phase HPLC, which separates peptides based on their hydrophobocities, is most commonly used in proteomics due to its compatibility and simplicity.

In this study, reverse phase HPLC is interfaced off-line with MALDI-FTMS. The eluent from the LC is directly deposited onto a multi-spot MALDI target that serves as a fraction collector. Capillary scale HPLC (150 μm i.d. x 15 cm columns) is used in this off-line technique since its low flow rate allows for the collection of typical MALDI sample volumes. A robotic autosampler can collect numerous fractions over the course of a HPLC separation. Following sample elution, MALDI matrix is applied to each sample spot by the autosampler.

While LC-ESI-MS is a very good technique for shotgun proteomic analysis, LC-MALDI-MS offers numerous advantages as well. The separation technique is decoupled from mass analysis. Deposition of the sample on a target plate provides a permanent record of the HPLC separation. A sample of interest can be re-examined since the analyte is embedded in the MALDI matrix. Peptides have different ionization efficiencies and since protein expression levels vary, their constituent peptides are present in varying abundances.[11] As a result, some peptides will produce intense peaks from a single MALDI event, while others require signal averaging (or ion accumulation in FT instruments) to obtain a reasonable mass spectrum. With on-line LC-ESI-MS, the sample is present transiently in the mass spectrometer as it elutes from the LC column. The entire separation procedure must be repeated in order to reexamine a particular peptide peak.

Another key advantage of MALDI is that it is far less prone to signal suppression during the analysis of complex mixtures than ESI. MALDI can tolerate many more peptide species per ionization event without masking peptides of low ionization efficiency or low abundance. Proteomic analysis by LC-ESI-MS tries to eliminate signal suppression through very high chromatographic resolution.[12] With high resolution, fewer peptide components are present per ionization event. However, long capillary columns and long gradient elution times are required to achieve such high resolution. With MALDI, such extreme chromatographic efforts are not necessary and wider fractions of peptides can be examined at one time. This greatly reduces the number of mass spectra that need to be acquired, stored, and analyzed.

An additional advantage of MALDI over ESI is that internal calibration using an external calibrant is easy to achieve using MALDI, but very difficult to achieve with ESI.[13,14] Internal calibration is required to obtain highly accurate mass measurements.[15] Finally, the data generated by MALDI and ESI is complementary in nature. Some peptides ionize more readily using one technique compared to the other. By combining data generated by both soft ionization methods, more overall proteins can be characterized.[16]

Accurate Mass Measurements Using LC-MALDI-FTMS. LC-MALDI-FTMS is an ideal technique for shotgun proteomic analysis. While MS/MS can be accomplished using FTMS, protein identification using accurate mass measurements of peptides is a method with the potential for higher-throughput.[17,18] In order to conclusively identify a protein from its constituent peptides, masses must be measured with a high degree of accuracy. When many peptides are present in a single mass spectrum, high resolution is also required to separate overlapping isotopic distributions. FTMS has unparalleled mass accuracy (~1 ppm) and mass resolution (routinely 100,000) as well as excellent sensitivity for peptides (low femtomole to attomole).[19] Furthermore, peptide identification cannot be relied upon if peaks that do not correspond to peptide molecular masses compromise the mass spectrum. Spurious peaks resulting from metastable decomposition of peptides has hindered peptide identification by MALDI-FTMS.[20]

MALDI under ultra-high vacuum can cause peptides to undergo metastable loss of one or more neutral molecules, usually water or ammonia.[20,21] Peaks resulting from these metastable losses are 17 or 18 mass units less than the monoisotopic peptide peak and can be quite abundant. For lower mass peptides, metastable loss peaks are usually small, i.e. 10% relative abundance or less. However, as peptide mass increases, metastable loss peaks increase in relative abundance and may even be mistaken for the molecular ion.[22] Metastable loss peaks will return spurious hits during database searching which reduces the reliability of protein identification.

The use of a “cool” matrix such as 2,5-dihydroxybenzoic acid (DHB) in MALDI analysis reduces the propensity for metastable decomposition. Nevertheless, metastable loss peaks from high mass peptides are of significant abundance. Infrared MALDI has also been known to reduce metastable decomposition.[23,24] Ions formed by infrared MALDI have less internal energy and hence more stable than those formed by ultraviolet MALDI.[25] Another method for suppressing metastable losses is the use of a high pressure MALDI source.[26,27] Metastable decomposition is eliminated at high pressures most likely as a result of collisional thermalization of ions.[26] The use of high pressure MALDI facilitates a shotgun approach to proteomics using LC-MALDI-FTMS.[11]

Methanococcus maripaludis. In this study, a select group of *Methanococcus maripaludis* proteins are analyzed using LC-MALDI-FTMS. *M. maripaludis* is a methane-producing archaeon, a prokaryotic organism distinct from bacteria in structure and chemistry.[28] Methanogens such as *M. maripaludis* are anaerobic and mainly inhabit the bottom of natural wetlands, rice paddies, sewage plants, and the digestive systems ruminants.[29] Methanogenesis, the sole source of ATP for methanogens, is the conversion of carbon dioxide and hydrogen to methane. A host of enzymes and coenzymes catalyze the process. Methanogens produce roughly twenty times more methane than all industrial sources combined and the combustion of methane yields the most energy per unit CO_2 produced of all hydrocarbon fuels.[30] Therefore, understanding the key components involved in methanogenesis is of great biotechnological value.

M. maripaludis is an ideal organism for a high-throughput proteomic study of methanogenesis. While most methanogens are slow growing extremophiles, *M. maripaludis* is a mesophile that grows rapidly using carbon dioxide as its sole carbon source.[31] In addition, isolation of proteins is assisted by the fact that detergents easily disrupt the cell envelope of *M. maripaludis*. Protein identification is aided by the fact that *M. maripaludis* is a relatively simple organism with a small genome. The genome has recently been sequenced and contains 1761 open reading frames (ORFs) that could encode proteins. Of these 1761 ORFs, 73% have been assigned functions based on DNA sequence homology with archaeal, bacterial, or eukaryotic proteins.[32]

Sub-20 kDa proteins of *M. maripaludis* were specifically chosen for this high-throughput proteomic study. First of all, many of these proteins have not been previously studied. Most proteomic studies to date have utilized 2D-GE, which has shown bias against low molecular weight proteins.[33] Additionally, many low molecular weight proteins have key roles in methanogenesis and its regulation. These proteins include ferredoxins and iron-sulfur proteins involved in electron transfer, ribosomal proteins involved in protein translational, and the individual enzymes responsible for methanogenesis. Of the 1761 ORFs of *M. maripaludis*, roughly one third code for sub-20 kDa proteins.

In this study, the utility of LC-intermediate pressure MALDI-FTMS as a technique for high-throughput proteomics is tested. A peptide mixture from a digested standard protein is first used to test the method for high mass accuracy, high mass resolution, and suppression of metastable losses. LC-IP MALDI-FTMS is then applied to analyze the sub-20 kDa proteins of *M. maripaludis* from a whole-cell lysate digest. Accurate measurements of peptide masses are used for protein identification.

EXPERIMENTAL

Sample Preparation. 1 mg/ml of the standard protein BSA, bovine serum albumin (Sigma-Aldrich, St. Louis, MO), was prepared using 100 mM ammonium bicarbonate. 250 μL of BSA was heat denatured at 95°C for 5 minutes. 12 μL of 50 mM TCEP, Tris(2-carboxyethyl) phosphine hydrochloride (Pierce, Rockford, IL), was added to reduce the disulfide bonds of BSA. The solution was thoroughly mixed and heated at 95°C for 10 minutes. Following cooling, 150 μL of 0.02 μg/μL sequencing grade modified trypsin (Promega, Madison. WI) was added to BSA. The solution was thoroughly mixed and digested for 16 hours at 37°C.

M. maripaludis proteins were harvested by the research group of Dr. William B. Whitman (Department of Microbiology, University of Georgia). Low molecular weight proteins were isolated using a Bio-Rad 491 preparative electrophoresis cell shown in Figure 3.1 (Hercules, CA). A 13% acrylamide gel was prepared, and 23.1 mg of crude extract from *M. maripaludis* was loaded onto the prep cell. 30 low molecular weight fractions were collected from the prep cell using electrophoresis running buffer (50 mM Tris, 0.384M glycine, and 0.2% SDS). SDS-PAGE of each fraction was performed, and it was determined that fractions 13-27 contained the sub-20 kDa proteins (Figure 3.2). Concentration of fractions 13-27 yielded 720 μL of 0.4 mg/mL sub-20 kDa proteins.

Buffer exchange using 3 kDa molecular weight cut-off (MWCO) filters was performed to concentrate the sample and reduce the concentrations of SDS, glycine, and Tris in the protein buffer solution. The sample was divided into two MWCO filters and 150 μL of 100 mM ammonium bicarbonate was added to each filter. The samples were centrifuged for 5 minutes and waste was removed. This process was repeated three more times resulting in ~1/40th dilution of the original contaminants. Following buffer exchange, 250 μL of ~1 mg/ml sub-20 kDa *M. maripaludis* proteins in 100 mM ammonium bicarbonate with 0.005% SDS, 0.009 M glycine, and 1.168 mM Tris was left. 125 μL of ~1mg/mL sub-20 kDa *M. maripaludis* proteins was heat denatured at 95°C for 5 minutes. 4 μL of 50 mM TCEP was added, and the solution was thoroughly mixed and heated at 95°C for 10 minutes. Following cooling, 75 μL of 0.02 μg/μL sequencing grade modified trypsin was added to *M. maripaludis* proteins. The solution was thoroughly mixed and digested for 16 hours at 37°C.

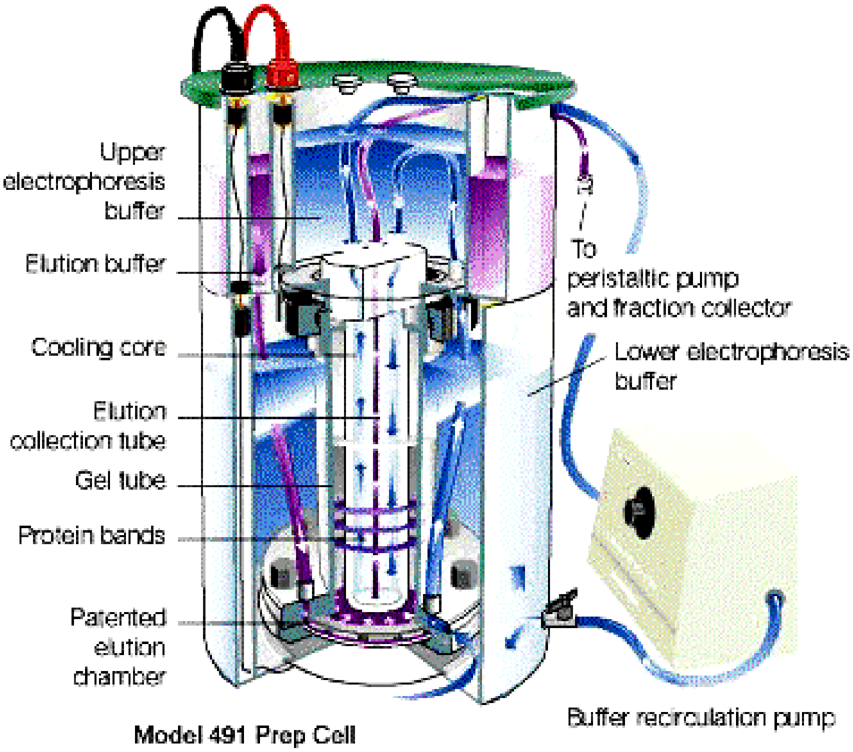

Figure 3.1 An illustration of a Bio-Rad 491 preparative electrophoresis cell. The image was extracted from http://www.bio-rad.com/electrophoresis/model491.htm.

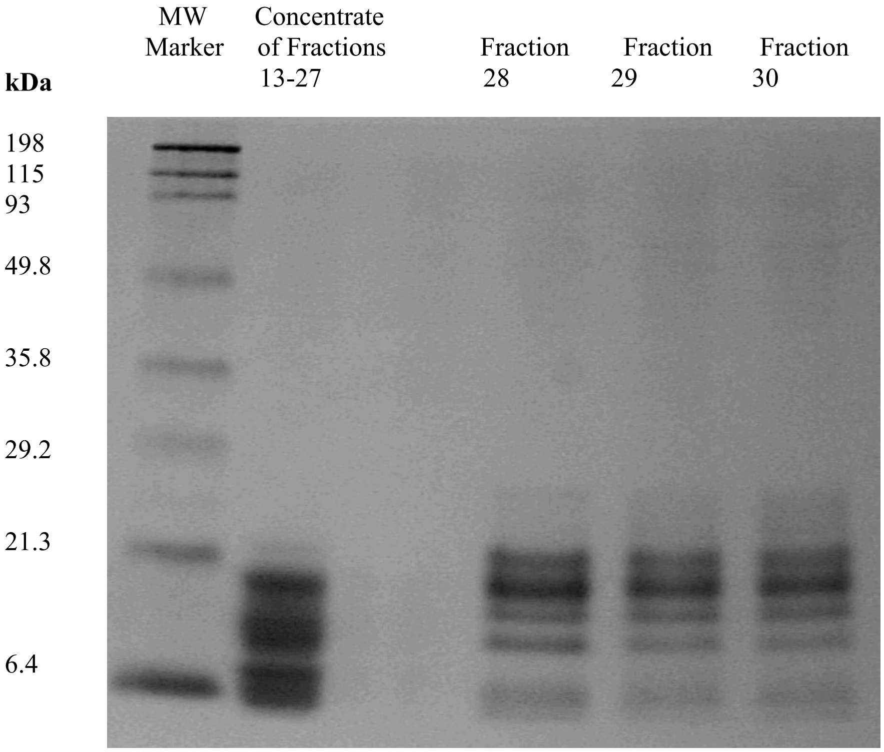

Figure 3.2 A photograph of a SDS-PAGE of the low molecular weight protein fractions of *Methanococcus maripaludis*.

Instrumentation. 5 μL of BSA tryptic digest was injected into a Micro-Tech Ultra Plus II MD capillary HPLC system (MicroTech, Vista, CA). Over the course of a 75-minute separation, 30 1 μL fractions were deposited on a multi-spot MALDI target using a Gilson 215 autosampler (Gilson, Middleton, WI). Following sample deposition, 1 μL of the recrystallized MALDI matrix DHB, 2,5-dihydroxybenzoic acid (Lancaster Synthesis, Pelham, NH), was applied to each sample spot by the autosampler and analyzed by intermediate pressure MALDI-FTICR MS using a 7.0 tesla BioApex (Bruker Daltonics, Inc., Billerica, MA).

5 μL of sub-20 kDa *M. maripaludis* digest was injected into a Micro-Tech Ultra Plus II MD capillary HPLC system. Over the course of a 75-minute separation, 30 1 μL fractions were deposited on a multi-spot MALDI target using a Gilson 215 autosampler. Following sample deposition, 1 μL of recrystallized DHB was applied to each sample spot by the autosampler. The samples were analyzed by intermediate pressure MALDI-FTICR-MS using a 7.0 tesla BioApex (Bruker Daltonics, Inc., Billerica, MA).

A 4.7 Tesla MALDI-FTICR-MS constructed in-house was used for vacuum MALDI experiments.[34] 1 μL of BSA digest and 1 μL of recrystallized DHB were manually deposited on a titanium MALDI target for analysis.

Data Analysis. BSA tryptic digest was analyzed using the MS-Fit program in Protein Prospector. Protein Prospector is a proteomics database search program developed by the Mass Spectrometry Facility at the University of California at San Francisco. An in-house database search program was used to analyze the *M. maripaludis* tryptic digest. The operation of each program is detailed in the Results and Discussion section.

RESULTS AND DISCUSSION

LC-IP MALDI-FTMS Analysis of a Standard Protein. The viability of LC-IP MALDI-FTMS as a method for high-throughput proteomics was tested using a standard protein digest. In order to conclusively identify a protein from its constituent peptides, the peptide masses must be measured at a high degree of mass accuracy (<10 ppm). The mass accuracy was examined using a BSA digest. 21 of the 30 LC fractions contained peptides and were analyzed using the Bruker software program XMASS. A combined 105 monoisotopic peptide peak masses were recorded. The peptide masses were compared with those in the SwissProt database using the MS-Fit program in Protein Prospector. The search was performed at a 10 ppm mass tolerance level, allowed for up to five missed cleavages by trypsin, and considered possible amino acid modifications including oxidation of methionine and the conversion of an N-terminal glutamine to pyro-glutamic acid. BSA was returned as the top result matching 32 of the 105 peptide masses for 42% sequence coverage. Table 3.1 lists the peptide masses matched to BSA. An average absolute error of 0.5 ppm was observed, with no error over 3.62 ppm. Since peptides were found in different mass spectra, relative abundances were calculated based upon the absolute abundances of each peptide. Roughly 20 of the 73 unmatched peptides masses corresponded to other bovine proteins possibly present in the BSA sample, including bovine lipoxygenase and bovine heavy chain myosin. Nonspecific cleavages can account for the remainder of the unmatched peptides. The majority of these peaks were of low relative abundance. Overall, the high degree of mass accuracy of the matched BSA peptides indicated that LC-IP MALDI-FTMS is suitable for peptide mass fingerprinting using accurate mass measurements.

Table 3.1 Analysis of BSA digest illustrating the high mass accuracy capability of LC-intermediate pressure MALDI-FTMS.

m/z Submitted	Error (ppm)	Relative Abundance	Amino Acid Sequence #	Missed Cleavages	Peptide Modifications
649.3339	-0.6485	0.029	205-209	0	
649.3339	-0.6485	0.029	223-228	0	
665.3769	-0.9115	0.114	156-160	0	
689.3732	-0.4116	0.102	236-241	0	
818.4253	-0.8222	0.018	562-568	0	
847.5044	0.2789	0.048	242-248	1	
906.4712	-0.7472	0.018	205-211	1	
922.4885	-0.0794	0.144	249-256	0	
927.4926	-1.5018	0.146	161-167	0	
994.3937	-0.3227	0.010	413-420	0	pyroGlu
997.5925	-0.8748	0.069	549-557	0	pyroGlu
1001.5860	-3.6189	0.139	233-241	1	
1014.6201	0.1757	0.074	549-557	0	
1142.7153	0.3633	0.053	548-557	1	
1163.6314	0.1592	0.222	66-75	0	
1249.6209	-0.6504	0.421	35-44	1	
1283.7108	-0.3088	0.032	361-371	0	
1305.7144	-1.7429	0.100	402-412	0	
1362.6724	-0.2657	0.081	89-100	0	
1399.6916	-1.1168	1.000	569-580	0	
1415.6870	-0.7613	0.034	569-580	0	1Met-ox
1439.8116	-0.4914	0.078	360-371	1	
1479.7961	0.0752	0.361	421-433	0	
1567.7433	0.0099	0.281	347-451	0	
1639.9339	-2.6806	0.052	437-451	1	
1667.8108	-1.7323	0.055	469-482	0	
1823.9017	0.8246	0.013	508-523	0	
1850.8972	-1.438	0.029	528-544	0	
1888.9341	3.5486	0.026	169-183	0	
2045.0275	-0.4929	0.160	168-183	1	
2199.1030	1.0248	0.045	562-580	1	
2441.1758	2.4215	0.017	524-544	0	

High mass resolution is also important for peptide mass fingerprinting using accurate mass measurements. Without high mass resolution, two or more peptide peaks may appear to be a single peptide peak. Therefore, to positively identify all components present in a complex sample, all monoisotopic masses must be resolved. An example of the resolving power of LC-IP MALDI-FTMS is found in a mass spectrum of a *M. maripaludis* digest fraction. Figure 3.3 illustrates a pair of baseline resolved peptide peaks. The monoisotopic peaks of each peptide are labeled, revealing only a 0.1381 mass unit difference. Among mass spectrometers, only FTMS is capable of this degree of mass resolution.

Suppression of metastable losses through the use of an intermediate pressure MALDI source was also tested in this study. A BSA tryptic digest was analyzed using both vacuum MALDI-FTMS and intermediate pressure MALDI-FTMS. Figure 3.4 shows a vacuum MALDI-FTICR mass spectrum of a BSA digest obtained with a 4.7 tesla instrument. Four separate metastable losses are seen in Figure 3.4. Even though the "cool" matrix DHB was used in analysis, metastable losses were still in excess of 25% the abundance of the peptide peak from which they derived. Suppression of metastable losses through use of an intermediate pressure MALDI source is illustrated in Figure 3.5. The large peak at 1567.7448 Da is a BSA peptide peak, while the very small peak at 1549.7345 Da is its metastable loss of water peak. The same peaks can be found in the vacuum MALDI spectrum. However, the metastable loss of water peak is <5% abundant using the intermediate pressure MALDI source compared to >25% abundant using the vacuum MALDI source. The intermediate pressure MALDI source is thus much more applicable to the analysis of complex peptide mixtures.

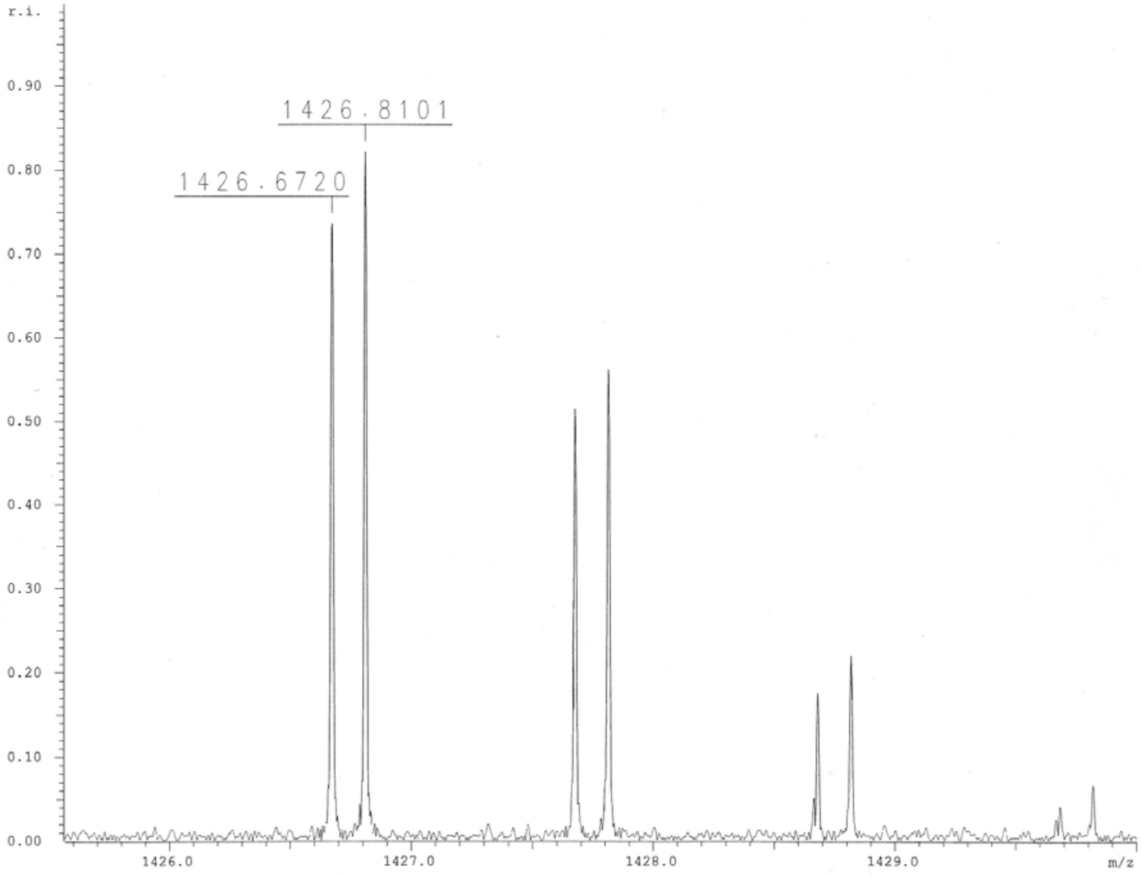

Figure 3.3 A mass spectrum of *Methanococcus maripaludis* peptides illustrating the high mass resolution capability of LC-intermediate pressure MALDI-FTMS. Two distinct monoisotopic peptide peaks are labeled at *m/z* 1426.6720 and *m/z* 1426.8101.

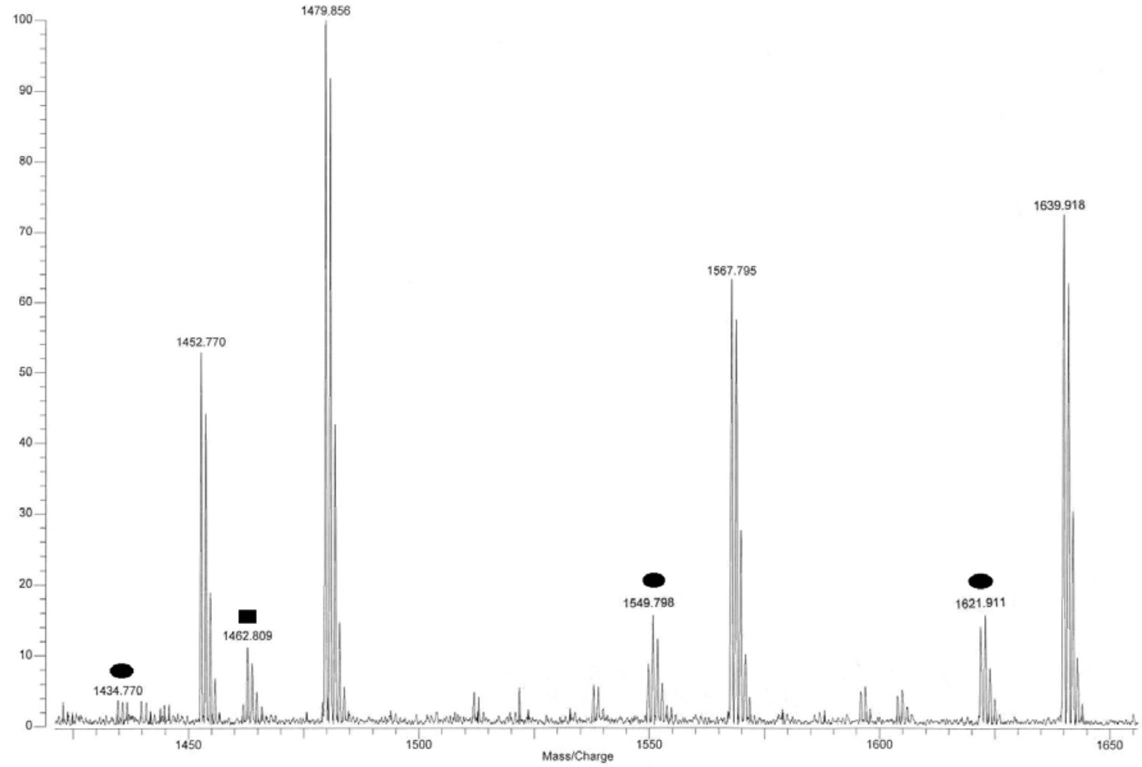

Figure 3.4 A mass spectrum of a BSA digest demonstrating metastable losses using vacuum MALDI-FTMS. ● = metastable peak from loss of water; ☐ = metastable peak from loss of ammonia

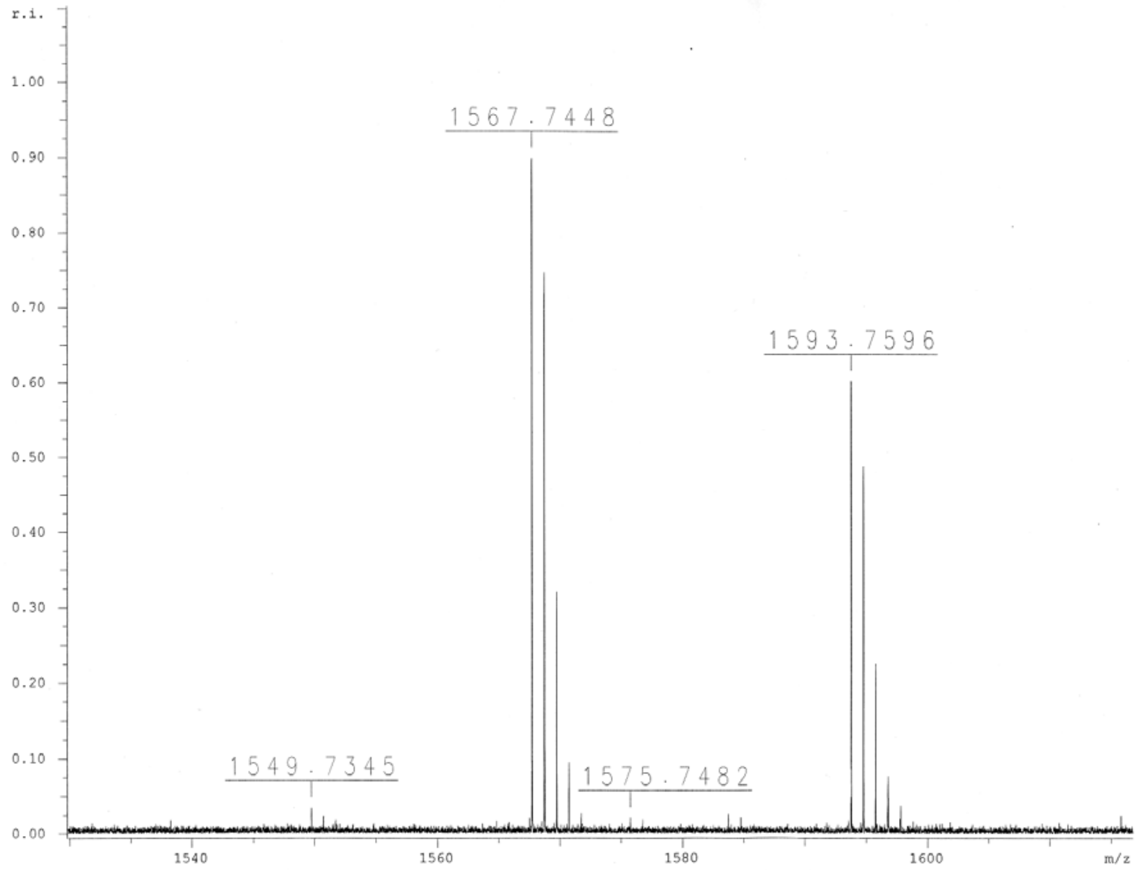

Figure 3.5 A mass spectrum of a BSA digest demonstrating suppression of metastable losses using LC-intermediate pressure MALDI-FTMS. The peaks at *m/z* 1549.7345 and *m/z* 1575.7482 are the result of metastable loss of water.

LC-IP MALDI-FTMS Analysis of Low Molecular Weight *Methanococcus maripaludis* Proteins. The LC-MALDI based proteomic method was then applied to a complex protein mixture – a sub-20 kDa protein fraction of *M. maripaludis.* 26 of the 30 LC fractions of a *M. maripaludis* tryptic digest contained peptide peaks and were analyzed using IP MALDI-FTMS. Figure 3.6 shows a two-dimensional contour plot of the entire LC run during which fractions were collected and analyzed. The horizontal axis corresponds to mass-to-charge ratio (*m/z*). The vertical axis is related to the LC elution time, with the first fraction collected beginning at 0.0 minutes. Each contoured spot indicates a peptide at a given *m/z* and relative elution time. Since reverse-phase HPLC separates peptides by hydrophobocity, early fractions (correlating to 0.0 to 0.4 minutes on the vertical axis) contain smaller, less hydrophobic peptides. Peptides over 2000 Da are not eluted until fraction 7 (0.54 - 0.62 minutes). As the elution gradient becomes less aqueous and more organic, larger and more hydrophobic peptides are eluted.

A horizontal mass spectral projection of all LC fractions is depicted at the top of Figure 3.6. The addition of all peptide peaks in the 26 LC fractions creates the mass spectral projection. A larger version of the *M. maripaludis* tryptic digest mass spectral projection is depicted in Figure 3.7. The mass spectrum is very complex with peptide peaks appearing at nearly every mass. Considerable overlap of peptide isotope peaks makes analysis of the mass spectrum virtually impossible. Figure 3.7 illustrates the necessity of chromatographic separation of complex peptide mixtures prior to mass spectral analysis. Using the Bruker software program XMASS, 1462 monoisotopic

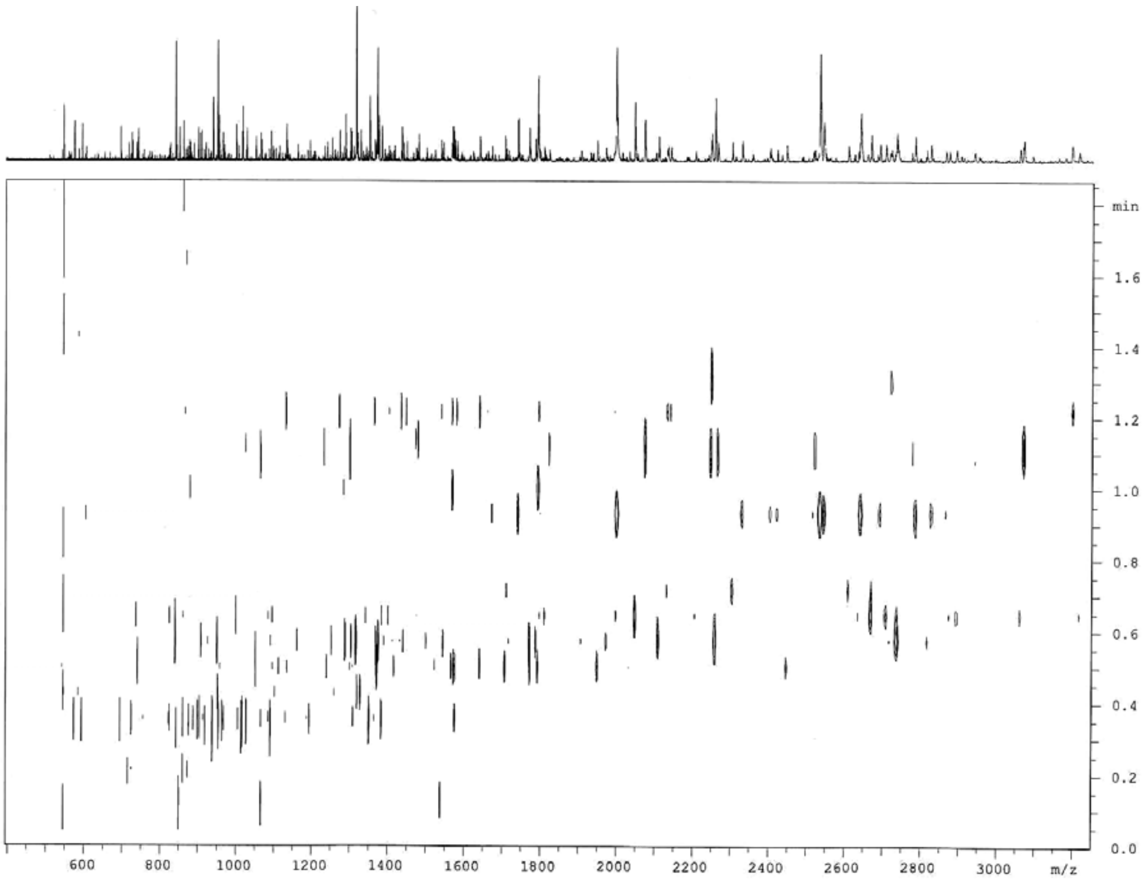

Figure 3.6 A two-dimensional contour plot and a horizontal mass spectral projection of a sub-20 kDa *Methanococcus maripaludis* protein tryptic digest produced by LC-intermediate pressure MALDI-FTMS.

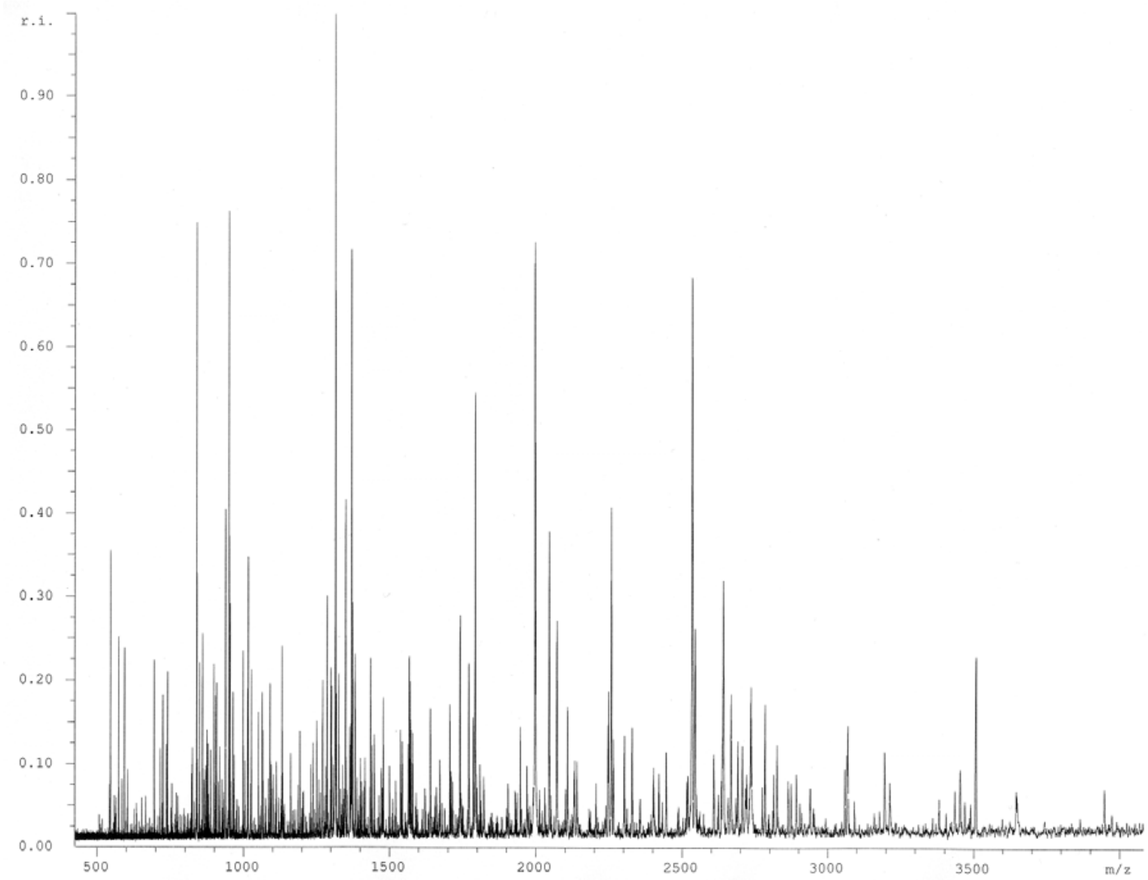

Figure 3.7 A combined mass spectrum of all LC fractions of a sub-20 kDa *Methanococcus maripaludis* protein tryptic digest analyzed by intermediate pressure MALDI-FTMS

peptide peaks were identified amongst the 26 mass spectra. Some of the 1462 peptide peaks represent the same peptide found in more than one fraction as a result of elution over more than one fraction.

Fractionation of the tryptic digest into individual MALDI spots simplifies analysis and allows for sample reinterrogation. An average of 55 monoisotopic peptide peaks are found in each LC fraction mass spectrum, with 150 peptides in the most abundant spectrum and 10 peptides in the least abundant spectrum. With such a limited number of peptide peaks, monoisotopic masses can easily be identified. This reduction in complexity is shown in Figure 3.8, a mass spectrum of LC fraction 2. Figure 3.8 illustrates the efficacy of reverse-phase HPLC in fractionating a complex peptide mixture. Only small, less hydrophobic peptides are eluted in fraction 2 since the elution gradient is more aqueous than organic. The majority of the 37 peptide peaks found in fraction 2 are under 1100 Da. Later eluting fractions exhibited more high mass peptides, as can be seen in the mass chromatogram shown in Figure 3.6.

Current commercial database search programs are not applicable to shotgun proteomics data. Search programs, such as MS-Fit in Protein Prospector, are designed to accept peptide masses from a single protein, not a complex mixture of proteins. An in-house database search program has been developed to analyze shotgun proteomics data. First, a list of all possible *M. maripaludis* tryptic peptides (allowing for up to two missed cleavages) was generated from the ORF database based on the genetic sequence of the organism. The 1462 monoisotopic peptide masses identified in the *M. maripaludis* mass spectra are compared against the masses in the list of possible tryptic fragments. A match

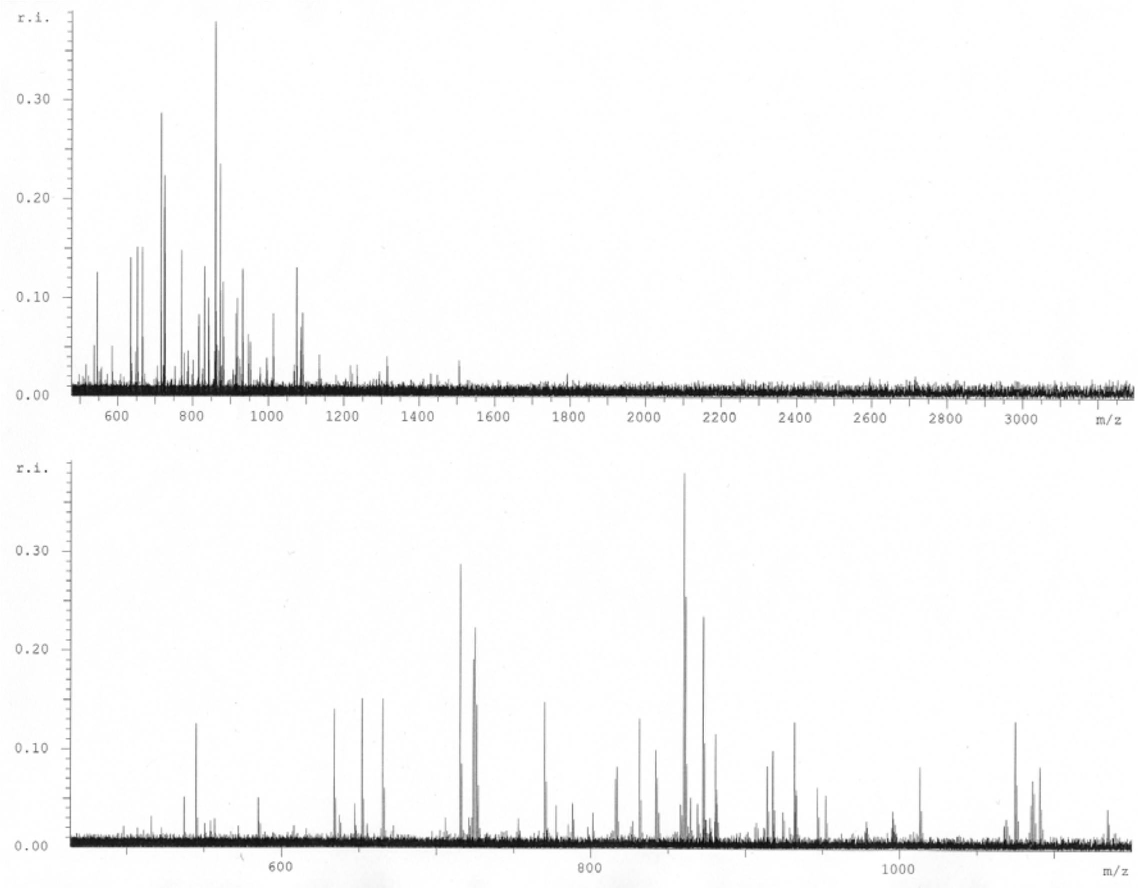

Figure 3.8 Mass spectra of LC Fraction 2 of a Sub-20 kDa *Methanococcus maripaludis* protein tryptic digest analyzed by intermediate pressure MALDI-FTMS. (Top) The complete mass spectrum from *m/z* 500-3000; (Bottom) Expansion of the mass spectrum from *m/z* 500-1150.

is recorded when an experimentally obtained peptide mass agrees with a calculated peptide mass within a specified mass accuracy.

The in-house search program first lists results according to experimentally obtained peptide mass. Possible tryptic peptides matching experimentally obtained masses within 5 ppm and 10 ppm are listed. A selected portion of the results is shown in Appendix A. Most of the 1462 monoisotopic peptide masses yielded numerous matches within both 5 ppm and 10 ppm. The search results were scanned manually for unique peptide masses. A peptide mass is considered unique if there is only one matching peptide within 10 ppm. Only 2% (29 of 1462) of the peptide masses were unique. 31% (9 of 29) of the unique peptides were from sub-20 kDa proteins. This percentage is equivalent to the percentage of ORFs that code for sub-20 kDa proteins in *M. maripaludis.* Only sub-20 kDa proteins were subjected to analysis in this study, so the number of unique peptides from sub-20 kDa proteins was expected to be greater. Poor mass calibration is a potential source of error in the experiment. However, even with <1 ppm mass calibration, many of the peptides will not be unique since different amino acids sequences can yield the same tryptic peptide mass, especially at lower masses (less than 1000 amu). Inclusion of proteins greater than 20 kDa is also a possibility for error since 9 additional unique peptides were from proteins with a molecular weight between 20-30 kDa. It is possible that larger proteins eluted with the sub-20 kDa fractions from the electrophoresis prep cell. Finally, errors in assigning reading frames in the DNA sequence data can produce errors in the protein database that would lead to incorrect molecular weight assignments for the proteins in the collected fraction.

Once peptide masses are assigned to a protein or to two or more candidates, the search program groups matches by protein. Proteins with experimental peptide masses that match actual peptide masses within 10 ppm are listed. The proteins are then ordered according to amino acid sequence coverage. Appendix B lists a selected portion of the results. It was found that nearly all proteins with greater than 40% sequence coverage were of low molecular weight. However, high sequence coverage is achieved for sub-20 kDa proteins with just a few peptides masses. 33% (3 of 9) of the unique peptides were matched to a sub-20 kDa protein with greater than 40% sequence coverage.

The experimentally obtained data was compared against a randomly generated data set for validation. A random data set of 1462 peptide masses was prepared using the list of all possible *M. maripaludis* tryptic peptides. The average mass of every nominal peptide mass from 700-2700 Da was calculated. A random number generator then selected 1462 of the averaged masses and indiscriminately altered each mass +/- 0.125 mass units to create the data set. The in-house search program analyzed the random data set and grouped matches by protein. The proteins were listed by sequence coverage. A histogram of the randomly generated data was prepared showing the molecular weight vs. sequence coverage distribution of the proteins (Figure 3.9). Tabular data of the histogram is presented in Table 3.2. A similar histogram was made using the experimentally obtained data (Figure 3.10), with tabular data present in Table 3.3. The two histograms show similar protein distributions. In addition, both histograms yield similar numbers of sub-20 kDa proteins with high sequence coverage. As can be seen, ordering protein matches by sequence coverage favors low molecular weight proteins. The data shows that increased specificity is required to positively identify proteins from

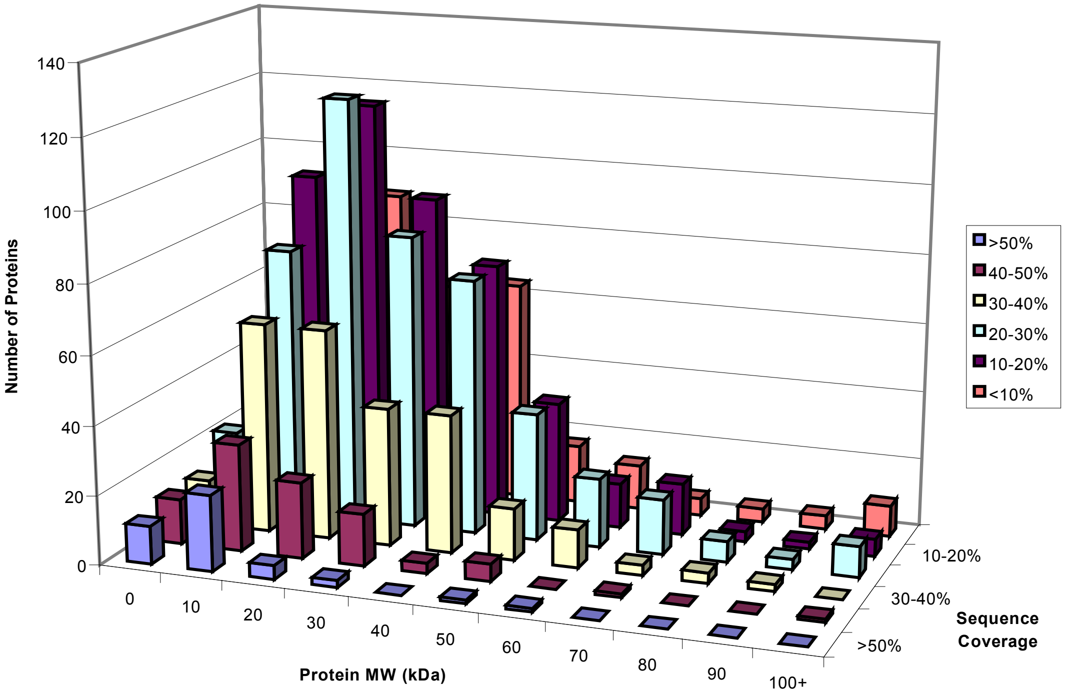

Figure 3.9 A histogram of protein molecular weight vs. sequence coverage distribution from 1462 randomly generated tryptic peptide masses of *Methanococcus maripaludis*.

Table 3.2 The number of proteins of a given molecular weight and sequence coverage range identified from 1462 randomly generated tryptic peptide masses of *Methanococcus maripaludis*.

	>50%	**40-50%**	**30-40%**	**20-30%**	**10-20%**	**<10%**
0-10 kDa	14	11	14	18	36	7
10-20 kDa	15	25	64	63	97	77
20-30 kDa	1	14	31	97	156	136
30-40 kDa	0	3	21	65	113	103
40-50 kDa	0	1	7	56	101	100
50-60 kDa	0	0	5	30	42	39
60-70 kDa	0	0	2	8	25	21
70-80 kDa	0	0	2	9	23	7
80-90 kDa	0	0	1	4	6	4
90-100 kDa	0	0	0	3	3	6
100+ kDa	0	0	1	3	10	7

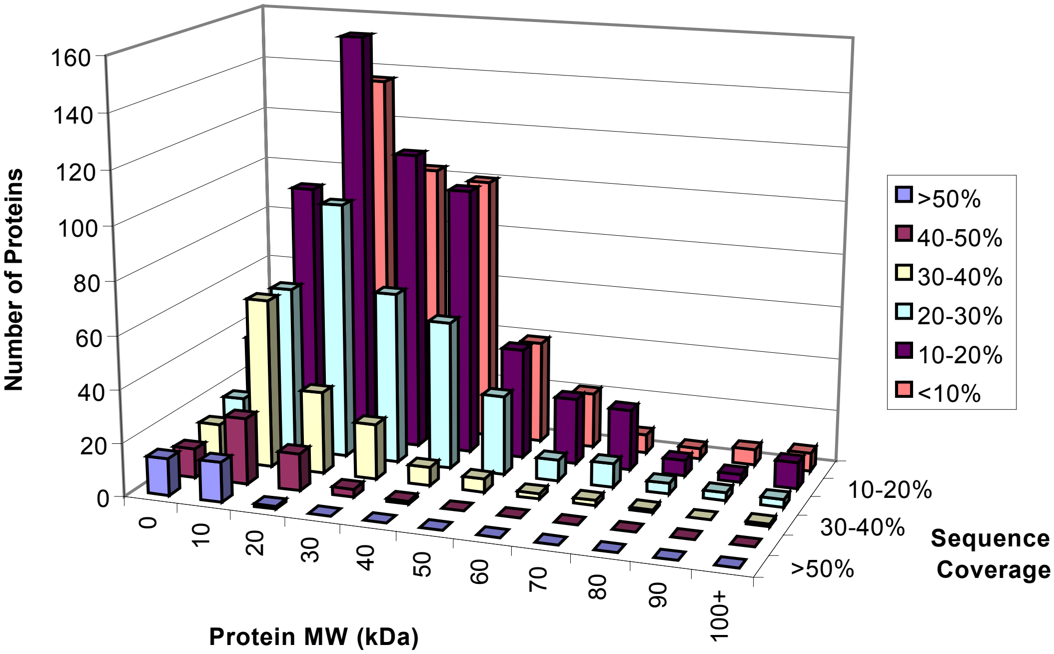

Figure 3.10 A histogram of protein molecular weight vs. sequence coverage distribution from 1462 experimentally obtained tryptic peptide masses of a sub-20 kDa *Methanococcus maripaludis* digest.

Table 3.3 The number of proteins of a given molecular weight and sequence coverage range identified from 1462 experimentally obtained tryptic peptide masses of a sub-20 kDa *Methanococcus maripaludis* digest.

	>50%	**40-50%**	**30-40%**	**20-30%**	**10-20%**	**<10%**
0-10 kDa	11	13	13	22	15	3
10-20 kDa	22	31	61	78	96	38
20-30 kDa	4	22	61	123	118	88
30-40 kDa	2	15	40	85	92	62
40-50 kDa	0	3	40	74	74	64
50-60 kDa	1	5	15	37	35	17
60-70 kDa	1	0	11	20	13	13
70-80 kDa	0	1	3	16	15	5
80-90 kDa	0	0	3	6	3	4
90-100 kDa	0	0	2	3	2	4
100+ kDa	0	1	0	9	5	9

the mass of their peptides. Such specificity can be obtained by measuring with higher mass accuracy, labeling specific amino acids, or using endogenous isotope labeling that permits one to establish the number of nitrogens or other elements in the peptides.

There are many possible factors contributing to the misidentification of proteins from a single peptide masses. Mass calibration is a potential source of error. The data were calibrated with an external standard. Internal calibration would provide better mass accuracy.[15] Without additional knowledge of the peptide, sub-1 ppm mass accuracy is required to identify a protein. Identification of a protein from accurate peptide masses relies heavily upon the validity of the protein database. Errors in the database from genetic sequencing miscues prevent proteins from being properly identified. Also, many proteins undergo post-translational modifications, such as glycosylation or phosphorylation. A protein cannot be correctly identified if the peptide mass has been altered by a post-translational modification. Nonspecific cleavages can also lead to erroneous matches since experimentally obtained data is matched only against possible tryptic peptide masses. Finally, outside contamination can lead to the misidentification of proteins. The peptide masses in the sample may not be from *M. maripaludis*. Stable-isotope labeling using ^{15}N-enriched media ensures that experimental peptide masses are from the organism of interest.

CONCLUSIONS

With high mass accuracy, high mass resolution, and the suppression of metastable losses, LC-intermediate pressure MALDI-FTMS is a high-quality technique for shotgun proteomics. In addition, the LC-MALDI based method provides rapid, high-throughput sample analysis with limited human intervention. Offline LC-MALDI creates a "permanent record" of the tryptic digests on the MALDI target allowing for re-examination of LC fractions. Analysis of the complex *M. maripaludis* peptide mixture yielded over 1400 monoisotopic peptide peaks using LC-IP MALDI-FTMS.

LC-MALDI-FTMS is adaptable to more complex samples. Employing a two-dimensional separation scheme with a strong cation exchange (SCX) fractionation prior to reverse phase HPLC would enable entire whole-cell lysates to be analyzed by MALDI-FTMS. In addition, special MALDI targets called anchor chips that consist of 400 μm hydrophilic spots arrayed on a hydrophobic surface can be used to improve the LC-MALDI method. Anchor chips condense sample deposition into a smaller area thereby concentrating the analyte. This improves sample sensitivity, reproducibility, and throughput.

While the performance of the LC-MALDI-FTMS method was excellent for batch protein digests, identification of the proteins using accurate mass measurements of peptides requires improvement. Even with 10 ppm mass accuracy, additional information about a peptide is required to positively identify a protein from its constituent peptides in complex mixtures. Our laboratory has recently proven that stable-isotope labeling of peptides using ^{15}N-enriched media assists in protein identification by eliciting the number of nitrogens in a given peptide. Also, the research presented in the following chapter involves the labeling of cysteine residues with a mass defect reagent. Labeled peptides can be identified based on their unique mass defect shift and the presence of a cysteine residue. Either of these techniques can be performed prior to LC-IP MALDI-FTMS to assist in protein identification using accurate mass measurements of peptides.

REFERENCES

1. Wolters, D. A., Washburn, M.P., and Yates, J.R. Anal. Chem. **2001,** 73, 5683.
2. Gygi, S. P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. *Nat. Biotech.* **1999,** *17*, 994.
3. Washburn, M. P., Wolters, D., and Yates, J.R. Nat. Biotech. **2001,** 19, 242.
4. Conrads, T. P., Alving, K., Veenstra, T.D. Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Pasa-Tolic, L., Udseth, H.R., Chrisler, W.B., Thrall, B.D., and Smith, R.D. *Anal. Chem.* **2001,** *73*, 2132.
5. Yao, X. D., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. Anal. Chem. **2001,** 73, 2836.
6. Shen, Y. F., Tolic, N., Zhao, R., Pasa-Tolic, L., Li, L.J., Berger, S.J., Harkewicz, R., Anderson, G.A., Belov, M.E., Smith, R.D. *Anal. Chem.* **2001,** *73*, 3011.
7. Smith, R. D., Pasa-Tolic, L., Lipton, M.S., Jensen, P.K., Anderson, G.A., Shen, Y.F., Conrads, T.P., Udseth, H.R., Harkewicz, R., Belov, M.E., Masselson, C., Veenstra, T.D. *Electrophoresis* **2001,** *22*, 1652.
8. Goodlett, D. R., Bruce, J.E., Anderson, G.A., Rist, B. Pasa-Tolic, L., Fiehn, O., Smith, R.D., and Aebersold, R. *Anal. Chem.* **2000,** *72*, 1112.
9. Griffin, T. J., Gygi, S.P., Rist, B., and Aebersold, R. Anal. Chem. **2001,** 73, 978-986.
10. Mortz, E., O'Connore, P.B., Roepstroff, P.Kelleher, N.L., Wood, T.D., McLafferty, F.W., and Mann, M. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 8264-8267.

11. Peters, E. C., Brock, A., Horn, D.M., Phung, Q.T., Ericson, C., Salomon, A.R., Ficarro, S.B., and Brill, L.M. *LC GC Eur.* **2002,** *15*, 423-428.

12. Shen, Y. F., Zhao, R., Belov, M.E., Conrads, T.P., Anderson, G.A., Tang, K.Q., Pasa-Tolic, L. Veenstra, T.D., Lipton, M.S., Udseth, H.R., Smith, R.D. *Anal. Chem.* **2001,** *73*, 1766.

13. Flora, J. W., Hannis, J.C., and Muddiman, D.C. *Anal. Chem.* **2001,** *73*, 5.

14. Hannis, J. C., and Muddiman, D.C. *J. Am. Soc. Mass Spectrom.* **2000,** *11*, 8.

15. Mize, T. H., and Amster, I.J. *Anal. Chem.* **2000,** *72*, 5586.

16. Arnott, D., O'Connell, K.L., King, K.L., and Shults, J.T. *Anal. Biochem.* **1998,** *258*, 1.

17. Lipton, M. S., Pasa-Tolic, L., Anderson, D.J., Auberry, D.L., Battista, K.R., Daly, M.J., Fredrickson, J., Hixson, K.K., Kostandarithes, H., Masselon, C., Markillie, L.M., Moore, R.J., Romine, M.F., Shen, Y.F., Stritmatter, E., Tolic, N., Udseth, H.R., Venkateswaran, A., Wong, L.K., Zhao, R., and Smith, R.D. *Proc. Natl. Acad. Sci. USA* **2002,** *99*, 11049-11054.

18. Smith, R. D., Anderson, G.A., Lipton, M.S., Pasa-Tolic, L., Shen, Y.F., Conrads, T.P., Veenstra, T.D., and Udseth, H.R. *Proteomics* **2002,** *2*, 513-523.

19. Marshall, A. G., Hendrickson, C.L., and Jackson, G.S. *Mass Spectrom. Revs.* **1998,** *17*, 1.

20. Castoro, J. A., and Wilkins, C.L. *Anal. Chem.* **1993,** *65*, 2621.

21. Li, Y. Z., Hunter, R.L., and McIver, R.T. *International Journal of Mass Spectrometry and Ion Processes* **1996,** *158*, 175.

22. McIver, R. T., Li, Y.Z., and Hunter, R.L. *Proc. Natl. Acad. Sci. USA* **1994,** *91*, 4801.

23. Zhang, W. Z., Niu, S.F., and Chait, B.T. *J. Am. Soc. Mass Spectrom.* **1998,** *9*, 879.

24. Niu, S. F., Zhang, W.Z., and Chait, B.T. *J. Am. Mass Spectrom.* **1998,** *9*, 1.

25. Budnick, B. A., Jensen, K.B., Jorgensen, T.D., Hasse, A., and Zubarev, R.A. *Rapid Commun. in Mass Spectrom.* **2000,** *14*, 578.

26. Laiko, V. V., Baldwin, M.A., and Burlingame, A.L. *Anal. Chem.* **2000,** *72*, 652.

27. O'Connor, P. B., and Costello, C.E. *Rapid Commun. in Mass Spectrom.* **2001,** *15*, 1862.

28. Howland, J. L. *The Surprising Archaea: Discovering Another Domain of Life*; Oxford University Press, Inc.: New York, 2000.

29. Doolittle, W. F. In *The Archaebacteria: Biochemistry and Biotechnology*; Portland Press: Chapel Hill, 1992, pp 1-6.

30. Large, P. J. *Methylotrophy and Methanogenesis*; Van Nostrand Reinhold Co. Ltd.: London, 1983.

31. Whitman, W. B., Sohn, S.H., Kuk, S.U., Xing, R.Y. *Appl. Environ. Microbiol.* **1987,** *53*, 2373-2378.

32. Badger, J. H., and Olsen, G.J. *Mol. Biol. Evol.* **1999,** *16*, 4233-4237.

33. Smith, R. D. *Nat. Biotech.* **2000,** *18*, 1041-1042.

34. Easterling, M. L., Pitsenberger, C.C., Kulkarni, S.S., Taylor, P.K., and Amster, I.J. *Int. J. Mass Spectrom. Ion Process.* **1996,** *158*, 97.

CHAPTER 4

INCREASING SPECIFICITY OF ACCURATE MASS MEASUREMENTS USING MASS DEFECT LABELING OF PEPTIDES

INTRODUCTION

In peptide mass fingerprinting, masses of peptides from proteins of interest are measured experimentally and compared with those predicted by the amino acid sequence in protein databases.[1] The more accurately masses are measured, and the more overall peptide masses that can be linked to the same protein, the greater the confidence in the final identification of the protein.[2,3] Increasing the specificity of a mass measurement or gaining additional information of the peptides being analyzed can also improve protein identification using accurate mass measurements of single peptides.[4] In this study, the viability of a novel mass defect reagent that increases mass measurement specificity by labeling cysteine-containing peptides is investigated.

Mass Defect Labeling. Examination of the mass distribution of all potential tryptic peptides reveals a clustering of masses at each nominal mass.[5] Peptides occupy narrow, predictable regions of the mass scale due to the small mass defects of their constituent elements (C, 0.0 mmu; H, 7.8 mmu; O, –5.1 mmu; N, 3.1 mmu; S, -28.0 mmu) and similar elemental compositions. Mass defect is the difference between the exact mass of a compound and the integer mass of the compound. Peptides exhibit a mass defect of approximately +0.06 amu per 100 amu of molecular weight. The positive mass defect of peptides mainly arises from the large fractional abundance of hydrogen atoms in a peptide molecule that each contributes +0.008 amu. Since peptides are large molecules with numerous atoms, they have significant mass defects. However, peptides only occupy a narrow range of a unit mass in the mass spectrum due to their similar mass defects. Figure 4.1 shows a histogram of all theoretical tryptic peptides (allowing up to 1 missed cleavage) between 1012-1014 Da of the archaeon *Methanococcus maripaludis*. The horizontal axis corresponds to molecular weight, with peptides in 5 ppm wide mass bins. The vertical axis represents the number of peptides in a mass bin. Peptide masses appear bunched in a ~0.3 Da mass range per unit mass resulting in significant overlap of peptide masses. Overlapping peptide masses reduce the specificity of an accurate mass measurement.

Greater mass measurement specificity can be achieved by spacing peptides over a larger portion of a unit mass. A novel derivitization procedure called mass defect labeling is designed to distribute peptides more evenly across the mass scale. Integration of heavy elements unique to the elemental composition of peptides will shift peptide masses to an unoccupied region of the mass scale. Bromine is an ideal heavy element to incorporate in a mass defect label since it has a large mass defect (-0.08 amu) and provides a unique mass spectral contribution due to its isotope distribution.

Cysteine residues are targeted for mass defect labeling because it has established alklyation chemistry and has a low frequency of occurrence in peptide sequences.[6] Labeling a frequently occurring amino acid would be counterproductive since the majority of peptides would be mass shifted. Figure 4.2 shows the mass defect label synthesized for this study, N-(2,4-dibromophenyl) maleimide. Labeling occurs as the free sulfhydryl of cysteine adds across the double bond of the maleimide portion of the molecule. This alklyation chemistry is illustrated in Figure 4.3 for the side-chain of a cysteine residue and N-ethylmaleimide.

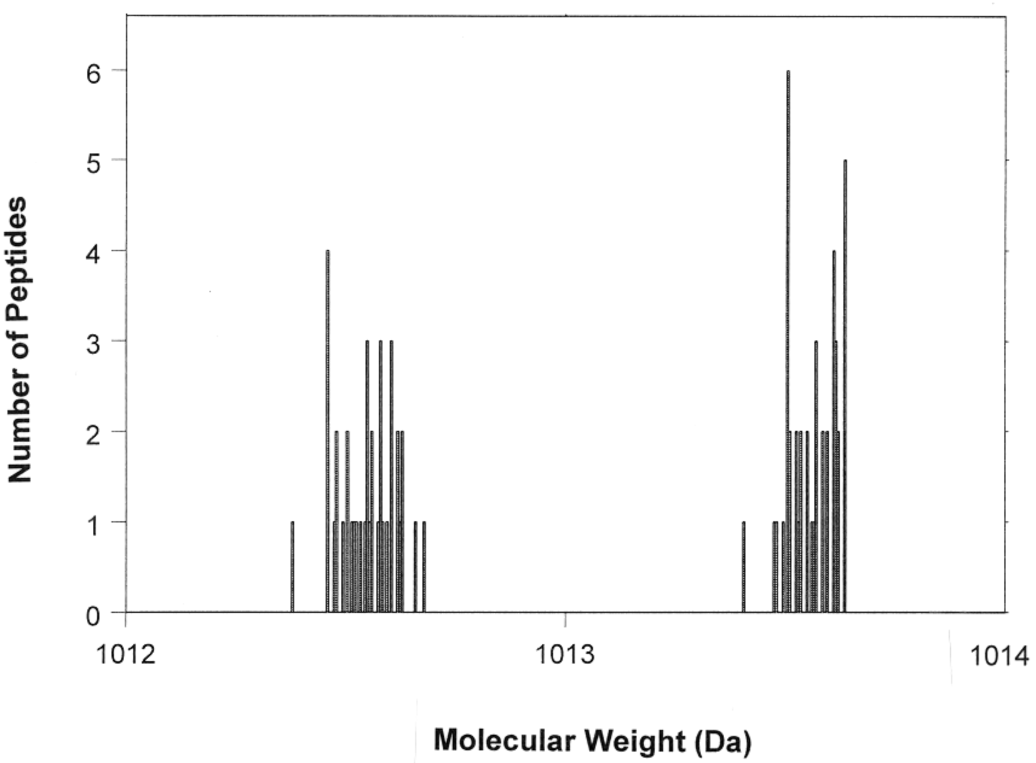

Figure 4.1 A histogram demonstrating the narrow distribution of theoretical tryptic peptide masses from *Methanococcus maripaludis*.

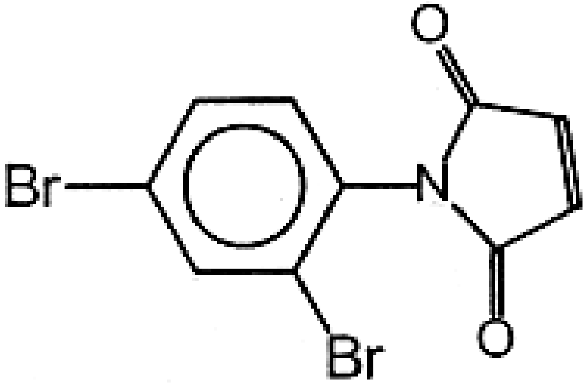

Figure 4.2 An illustration of the mass defect labeling reagent N-(2,4-dibromophenyl) maleimide.

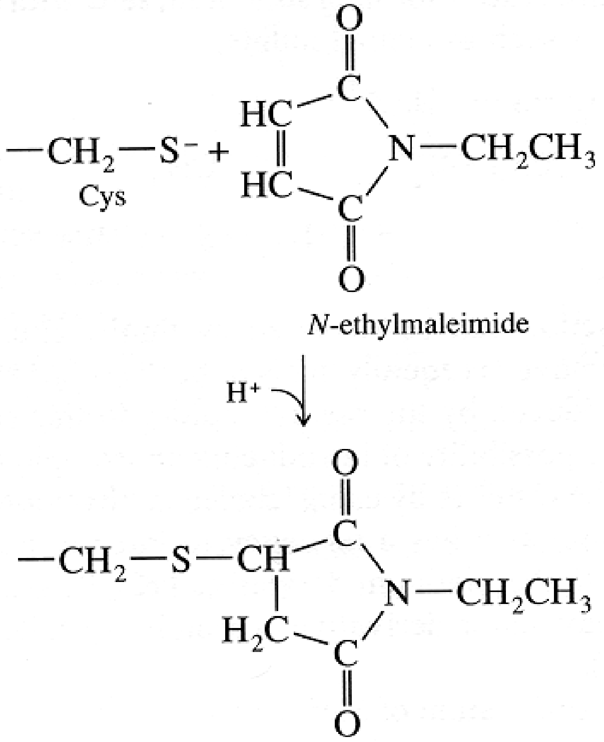

Figure 4.3 A schematic showing the addition of N-ethylmaleimide to the free sulfhydryl of a cysteine side-chain. The image was adapted from reference 6.

The effect of labeling peptides with N-(2,4-dibromophenyl) maleimide is easy to predict. Figure 4.4 shows the distribution of mass defects for all tryptic peptides of *M. maripaludis* between 700 and 3000 Da (allowing up to 2 missed cleavages), in which all cysteine residues have been labeled with N-(2,4-dibromophenyl) maleimide. The horizontal axis corresponds to the mass difference between the calculated mass defect and the average mass for all peptides of the same integer mass, with peptide masses in 10 ppm wide mass bins. The vertical axis is the number of peptides in a mass difference bin. Since the majority of tryptic peptides do not contain cysteine, the large central distribution of masses corresponds to unlabeled peptides. The majority of these unlabeled peptides have mass defects that lie within roughly +/- 0.15 amu of the main distribution. Peptides with one labeled cysteine residue are located 0.3 amu below the central distribution. This mass defect shift arises from two components. First, the dibromo-labeling reagent adds a –0.1 amu mass defect to the peptide. Secondly, the label increases the peptide mass by 328 amu placing it on a mass scale next to unlabeled peptides with larger mass defects. Since the mass defect of peptides is +0.06 amu per 100 amu of molecular weight, the unlabeled peptides will have a molecular mass of 0.197 amu larger than the peptide portion of the derivitized peptide. These two factors result in the overall -0.3 amu mass defect shift between labeled and unlabeled peptides. Peptides with two labeled cysteines will be shifted -0.6 amu. These peptides are illustrated in Figure 4.4 by the very small distribution at 0.4 amu, which correlates to a –0.6 mass shift on a 1 amu wide mass scale.

The utility of mass defect labeling for shotgun proteomics using accurate mass measurements can be modeled using the proteome of *M. maripaludis*. First, a list of all

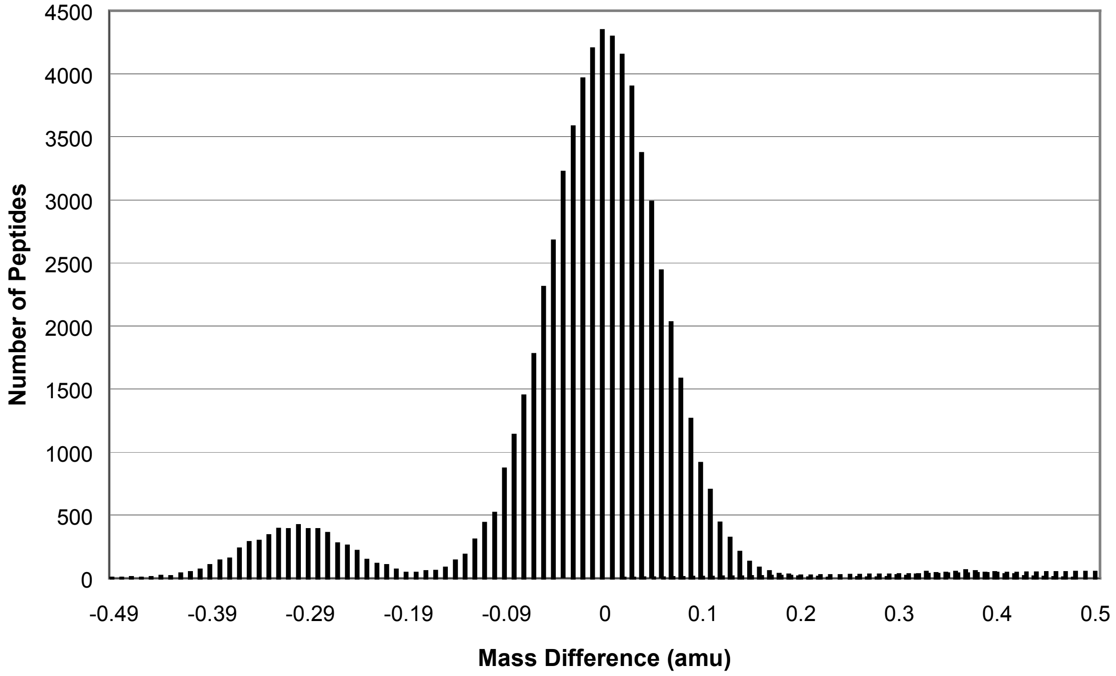

Figure 4.4 A histogram demonstrating the mass defect distribution of *Methanococcus maripaludis* tryptic peptides. The cysteine residues are derivitized by N-(2,4-dibromophenyl) maleimide.

possible tryptic peptides (allowing up to 1 missed cleavage) is calculated from the protein sequences of the archaeon. The tryptic masses are then sorted into 10 ppm wide bins (corresponding to +/-5 ppm mass accuracy). Each bin is then inspected to determine the percentage of bins that contain only a single peptide mass. These peptide masses are considered unique. Only 29% of the unlabeled peptide masses are unique when measured at 5 ppm mass accuracy. However, 77% of the mass defect labeled peptides can be uniquely assigned to one protein. Since 18% of all *M. maripaludis* peptides are cysteine-containing peptides, 43% of all peptides (29% unlabeled and 77% labeled peptides) can be assigned to a single protein based solely on one accurate mass measurement. With greater mass accuracy, these percentages increase dramatically. At 1 ppm mass accuracy which can be achieved by FTMS, 57% of unlabeled peptides are of unique mass, while 91% of the labeled peptides are unique to one protein. This allows up to 73% of *M. maripaludis* tryptic peptides to be assigned from a single accurate measurement of their molecular weight.

There are many benefits of mass defect labeling. Unlike ICAT reagents that use affinity tags to label cysteine-containing peptides, labeled and unlabeled peptides are not separated prior to MS analysis in mass defect labeling. This saves time by eliminating a step in the analysis process, and more importantly, peptides are not lost in a complicated separation procedure. Mass defect labeling allows both labeled and unlabeled peptides to be analyzed simultaneously. Therefore, database searches are not constrained to only peptides with a specific amino acid. Typical ICAT identification methods limit the searchable database to only cysteine-containing peptides. Misidentification of peptides can then arise since unlabeled peptides may be present in the labeled mixture. Mass defect labeling searches the entire database, using the presence of a cysteine only as a method to confirm protein identification. Additionally, mass defect labeling increases search specificity by decongesting the mass spectrum. By shifting labeled peptides to previously unoccupied mass space, the specificity of accurate mass measurements increases.

EXPERIMENTAL

Synthesis of Mass Defect Label. The mass defect label, N-(2,4-dibromophenyl) maleimide, was synthesized by an undergraduate student in our laboratory in a two-step process. In step 1, N-(2,4-dibromophenyl) maleanilic acid was prepared from 0.989 g maleic acid in 13.0 mL ether and 2.512 g dibromoaniline in 4.0 mL ethyl acetate. The mixture was stirred at room temperature for 1 hour and then cooled to 12°C. The product was collected, rinsed in ether, and dried via vacuum filtration resulting in 1.435 g (41.5% yield) of N-(2,4-dibromophenyl) maleanilic acid. In step 2, N-(2,4-dibromophenyl) maleimide was prepared from 0.328 g anhydrous sodium acetate in 3.4 mL acetic anhydride and 1.435 g N-(2,4-dibromophenyl) maleanilic acid. Following a brief heating to dissolve the mixture, 6.5 mL of ice water was added to the mixture resulting in an aqueous layer and an organic layer. The layers were separated and the organic layer was refrigerated for two days to induce crystallization. The product was then collected and dried by vacuum filtration. The product was recrystallized using cyclohexane and washed numerous times using ice water and petroleum ether to ensure purity. 0.064 g (4.7% yield) of N-(2,4-dibromophenyl) maleimide was produced and the structure was confirmed by mass spectrometry.

Sample Preparation. 1 mg of laminin fragment peptide (Cys-Asp-Pro-Gly-Tyr-Ile-Gly-Ser-Arg) was obtained from Sigma-Aldrich (St. Louis, MO). Laminin was suspended in 1 mL of 100 mM ammonium bicarbonate and diluted to 0.5 mg/mL. The mass defect label, N-(2,4-dibromophenyl) maleimide, was prepared at 20 mg/mL in methanol. 50 mM TCEP, Tris(2-carboxyethyl) phosphine hydrochloride (Pierce, Rockford, IL), was prepared in nano-pure water. 100 μL of 0.5 mg/mL laminin was distributed to five Eppendorf tubes to explore different experimental variables – (1) laminin (control); (2) laminin and 1-fold molar excess mass defect label; (3) laminin, 1-fold molar excess mass defect label, and TCEP; (4) laminin and 5-fold molar excess mass defect label; (5) laminin, 5-fold molar excess mass defect label, and TCEP. 5 μL of 50 mM TCEP was added to appropriate samples, which were then heated for 10 minutes at 95 °C. 0.85 μL of 20 mg/mL N-(2,4-dibromophenyl) maleimide was added as a 1-fold molar excess to appropriate samples, while 4.25 μL of 20 mg/mL N-(2,4-dibromophenyl) maleimide was added as a 5-fold molar excess to appropriate samples. Samples incubated at room temperature for 1 hour with occasional mixing.

Instrumentation. A 4.7 tesla MALDI-FTICR-MS constructed in-house was used in the mass defect labeling experiments.[7] 1 μL of each laminin sample and 1 μL of recrystallized DHB, 2,5-dihydroxybenzoic acid (Lancaster Synthesis, Pelham, NH), were manually deposited on a titanium MALDI target for analysis.

RESULTS AND DISCUSSION

A laminin fragment containing nine amino acid residues, one of which is cysteine, was analyzed in this mass defect labeling study. The computer program IsoPro was used to model the mass spectra of unlabeled laminin and laminin labeled with the mass defect reagent. The mass spectrum of unlabeled laminin has an isotope distribution of a typical peptide with a monoisotopic base peak at *m/z* 967.431 (Figure 4.5A). MALDI-FTMS analysis of the unlabeled laminin sample is identical to that predicted by IsoPro with a monoisotopic base peak at *m/z* 967.434 (Figure 4.5B).

Figure 4.6A demonstrates that the simulated mass spectrum of laminin labeled with the mass defect reagent is very different from the unlabeled laminin mass spectrum. The monoisotopic peak of labeled laminin is now found at *m/z* 1296.299 and is no longer the base peak. In addition, a mass defect shift of the peptide is observed since the monoisotopic peak of labeled laminin is found ~0.3 mass units lower than unlabeled peptides of *m/z* 1296. The unique mass spectrum arises from the fact that the mass defect reagent contains two bromine atoms. There are two naturally occurring isotopes of bromine, ^{79}Br and ^{81}Br that are present in abundances of 0.507 and 0.493 respectively. Bromine is not normally found in any amino acid residues; therefore peptides labeled with the mass defect reagent have a distinctive isotope pattern that is easily discernible by mass spectrometry. Since the mass defect label reagent is highly specific to cysteine residues, only cysteine-containing peptides will have a unique isotopic signature and mass defect shift.

MALDI-FTMS analysis of the laminin samples labeled with the mass defect reagent yielded a mass spectrum matching that predicted by IsoPro (Figure 4.6B). Laminin samples that contained the mass defect labeling reagent and TCEP gave similar mass spectra. Therefore, the presence of TCEP did not interfere with the labeling process. TCEP is commonly added to protein digests to reduce disulfide bonds formed

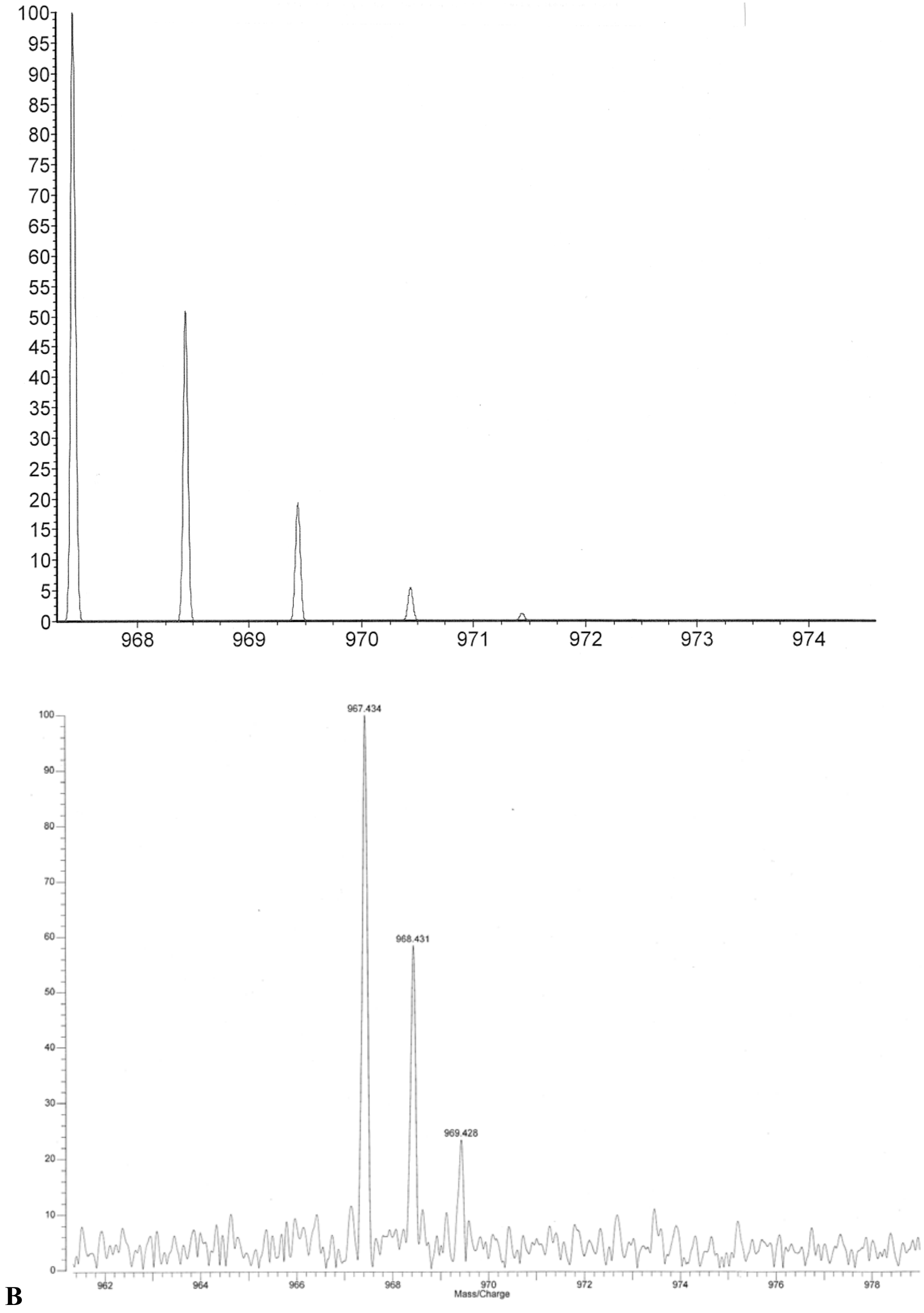

Figure 4.5 The modeled isotopic pattern of laminin peptide in a mass spectrum (A); MALDI-FTICR mass spectrum of laminin peptide (B).

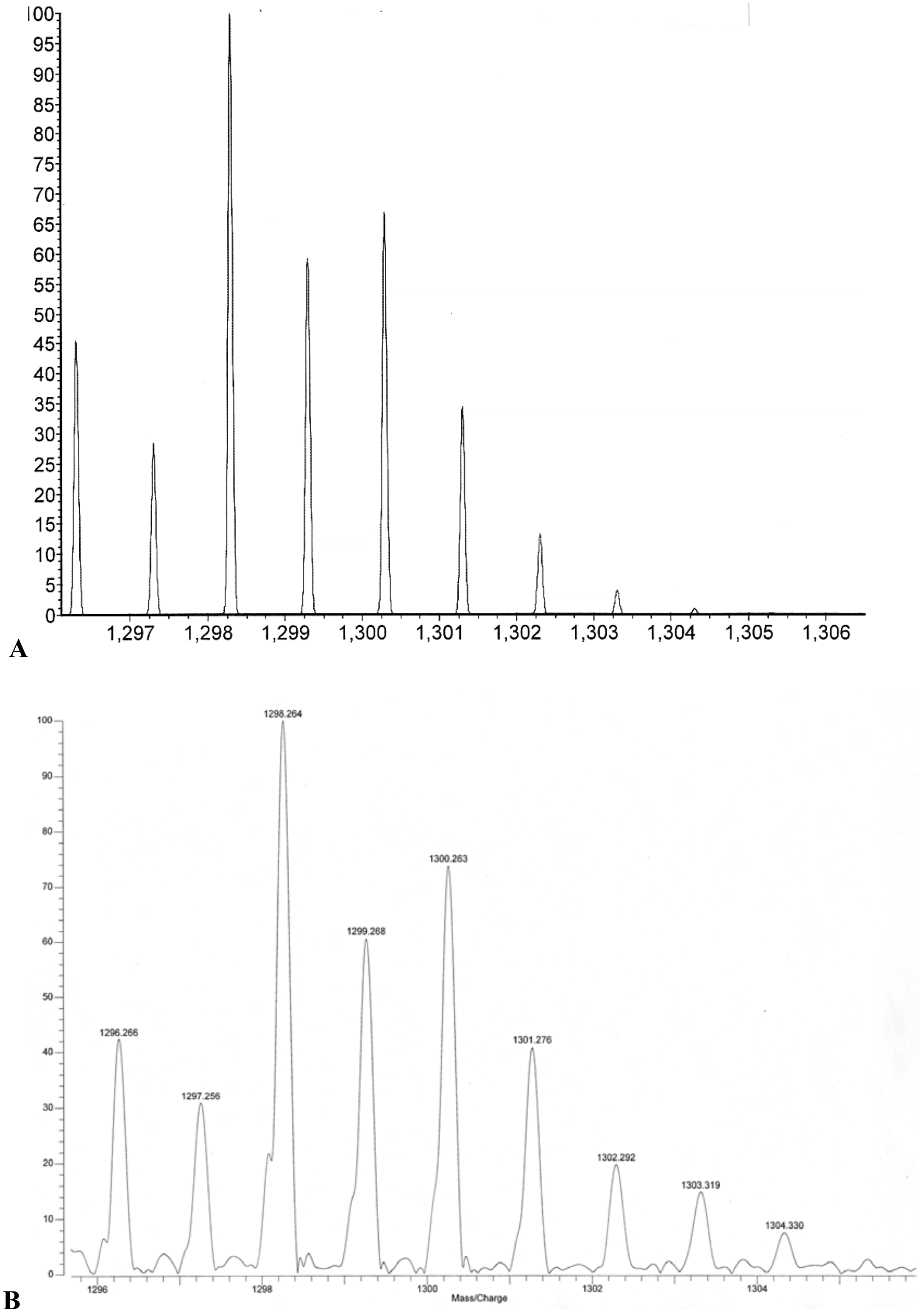

Figure 4.6 The modeled isotopic pattern of laminin peptide labeled with mass defect reagent in a mass spectrum (A); and MALDI-FTICR mass spectrum of laminin peptide labeled with mass defect reagent (B).

by neighboring cysteine residues to free sulfhydryls allowing them to be labeled. Additionally, it was determined that an equal molar addition of mass defect reagent was not enough to completely label laminin. A mass spectrum of laminin with a 1-fold molar excess of mass defect reagent, shows peptide peaks of both unlabeled and labeled laminin (Figure 4.7). Note that the largest peak for each peptide is labeled in the mass spectrum, *m/z* 967.419 for unlabeled laminin and *m/z* 1298.288 for labeled laminin. The smaller monoisotopic peak of labeled laminin is actually found two mass units lower at *m/z* 1296.289. In contrast to Figure 4.7, when laminin is labeled with a 5-fold molar excess of mass defect reagent, only labeled laminin peptides are evident in the mass spectrum (Figure 4.8).

CONCLUSIONS

In this study, the effectiveness of the mass defect label N-(2,4-dibromophenyl) maleimide was confirmed using the standard peptide laminin. The predicted mass defect shift and the unique isotope distribution of a peptide labeled with a dibromo-labeling reagent were observed. The mass defect labeling process is irreversible and unaffected by MALDI. It was also concluded that the reducing reagent, TCEP, did not affect the labeling of cysteine residues, and the addition of a 5-fold molar excess of the mass defect reagent resulted in complete labeling of the laminin peptide. The mass defect labeling process can now be applied with confidence to a standard protein digest and then to a whole-cell lysate using LC-intermediate pressure MALDI-FTMS for analysis.

Mass defect labeling is not restricted to cysteine-containing peptides. Using established alklyation chemistry, mass defect reagents can be synthesized for other

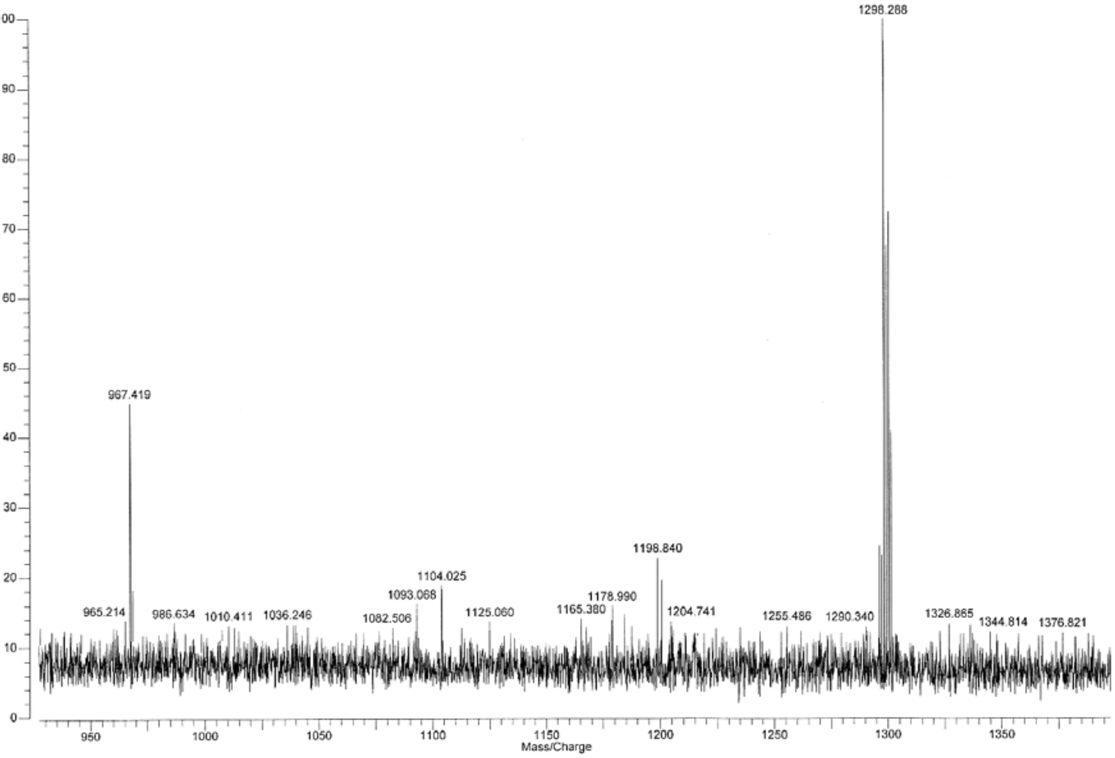

Figure 4.7 A MALDI-FTICR mass spectrum of laminin peptide labeled with 1-fold molar excess mass defect reagent.

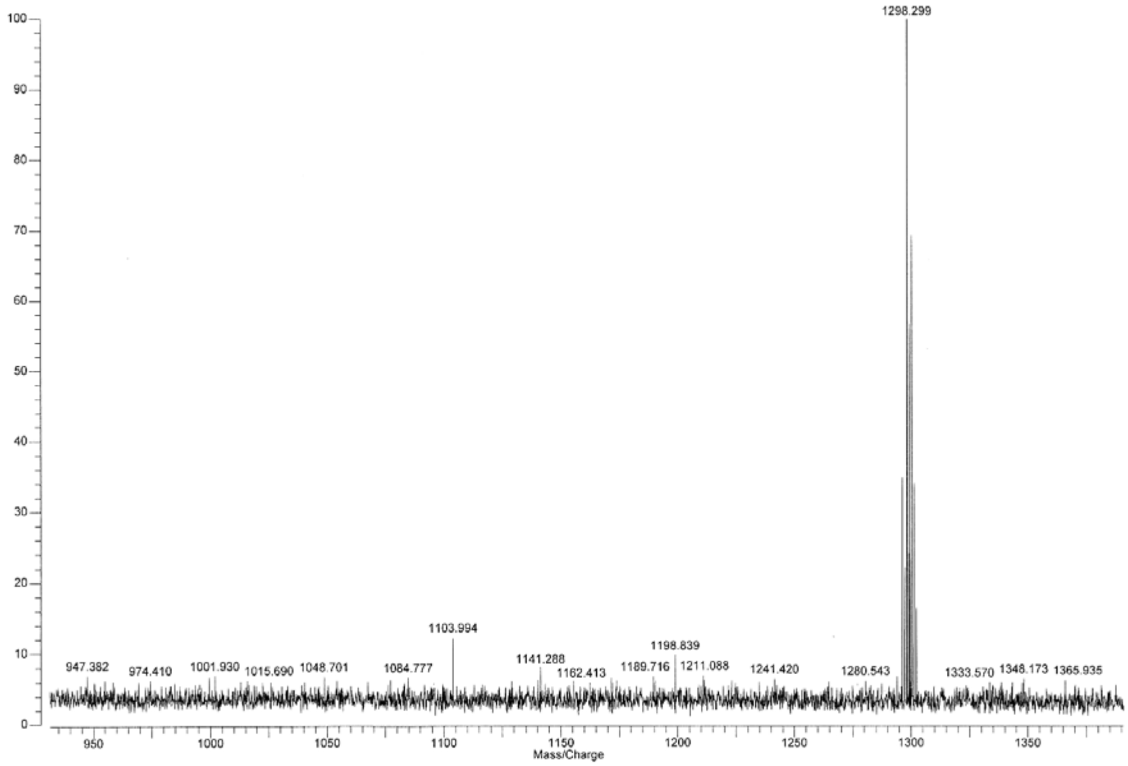

Figure 4.8 A MALDI-FTICR mass spectrum of laminin peptide labeled with 5-fold molar excess mass defect reagent.

infrequently occurring amino acids such as histidine, tryptophan, and methionine. Mass defect labeling is amenable to other types of proteomic studies as well. Quantitative proteomics is feasible using mass defect labeling. Stable-isotope labeling of peptides is possible given the synthesis of 100% ^{79}Br and 100% ^{81}Br mass defect labeling reagent. Each mass defect label contains two bromine atoms; therefore a 4 amu mass difference will be seen between peptides labeled with the light reagent (^{79}Br) and those labeled with the heavy reagent (^{81}Br). Mass defect labeling can also be used in conjunction with endogenous stable-isotope labeling. ^{15}N-metabolic labeling of peptides prior to mass defect labeling will allow for even greater protein identification and quantitation specificity.

REFERENCES

1. Mortz, E., Vorm, O., Mann, M., and Roepstorff, P. *Biol. Mass Spectrom.* **1994,** *23*, 249.
2. Fenyo, D., Qin, J., and Chait, B.T. *Electrophoresis* **1998,** *19*, 998-1005.
3. Zubarev, R. A., Hakansson, P., and Sundqvist, B. *Anal. Chem.* **1996**, 4060-4063.
4. Goodlett, D. R., Bruce, J.E., Anderson, G.A., Rist, B. Pasa-Tolic, L., Fiehn, O., Smith, R.D., and Aebersold, R. *Anal. Chem.* **2000,** *72*, 1112.
5. Hughey, C. A., Hendrickson, C.L., Rodgers, R.P., and Marshall, A.G. *Anal. Chem.* **2001,** *73*, 4676.
6. Creighton, T. E. *Proteins, Structures and Molecular Properties*; Freeman: New York, 1993.
7. Easterling, M. L., Pitsenberger, C.C., Kulkarni, S.S., Taylor, P.K., and Amster, I.J. *Int. J. Mass Spectrom. Ion Process.* **1996,** *158*, 97.

CHAPTER 5

CONCLUSIONS

In recent years, proteomics research has been directed towards the development of viable alternatives to two-dimensional gel electrophoresis (2D-GE).[1-8] The majority of current research involves the rapid analysis of entire proteomes through batch digestion of whole-cell lysates or cell sub-fractions. [1-7] This approach, known as shotgun proteomics, produces a complex mixture of peptides. Proteins are then identified from their constituent peptides. The research detailed in this thesis used MALDI-FTMS to obtain accurate mass measurements of peptides for protein identification.

An LC-MALDI-FTMS method for rapid, high-throughput proteomics was presented in Chapter 3. The method provided excellent mass accuracy and resolution while suppressing metastable losses through the use of an intermediate pressure MALDI source. Analysis of a complex biological sample, the low molecular weight proteins of *Methanococcus maripaludis,* yielded over 1400 monoisotopic peptide peaks. An in-house database search program indicated that protein identification from a single accurate mass measurement of a peptide was not a feasible high-throughput technique. Increased specificity of the accurate mass measurement or additional information about the peptide being analyzed is required for the identification of numerous proteins.

In Chapter 4, a technique for increasing the specificity of accurate mass measurements was described. Mass defect labeling of cysteine-containing peptides using N-(2,4-dibromophenyl) maleimide shifts the mass defect of the peptides –0.3 amu to a previously unoccupied region of the mass scale. Since cysteines are a rarely occurring amino acid, an accurate mass measurement of labeled peptides is highly specific. In addition, the presence of two bromine atoms in the mass defect label produces a unique isotope distribution for cysteine-containing peptides. Labeling of the laminin fragment peptide with a 5-fold molar excess of the mass defect reagent was successful. Mass defect labeling can now be applied with confidence to more complex samples.

Future high-throughput proteomic studies utilizing mass defect labeling for analysis of complex biological samples look promising. Mass defect labeling significantly increases the number of unique peptides in a composite mixture allowing for numerous proteins to be identified from single peptide mass measurements. Further peptide specificity can be achieved by using mass defect labeling in conjunction with endogenous stable-isotope labeling. Our laboratory has successfully identified *Methanococcus maripaludis* proteins from accurate mass measurements of ^{15}N-metabolic labeled peptides. Using ^{15}N-enriched media, the proteins of archaea or bacteria can be endogenously labeled and isolated. The proteins can then be exogenously labeled with the mass defect reagent prior to LC-MALDI-FTMS analysis.

Quantitative proteomic studies using mass defect labeling are also possible with or without endogenous stable-isotope labeling. Both light (100% ^{79}Br) and heavy (100% ^{81}Br) versions of the mass defect label can be synthesized. The light mass defect reagent labels a control group of proteins, while the heavy reagent labels a stressed group of the proteins. The proteins are then pooled and analyzed by LC-MALDI-FTMS. Since each mass defect label contains two bromine atoms, light and heavy peptide pairs are found 4 amu apart in the mass spectrum. Up- or down-regulation of proteins can be determined from the relative abundances of peptide pairs.

REFERENCES

1. Wolters, D. A., Washburn, M.P., and Yates, J.R. *Anal. Chem.* **2001,** *73*, 5683.
2. Washburn, M. P., Wolters, D., and Yates, J.R. *Nat. Biotech.* **2001,** *19*, 242.
3. Gygi, S. P., Rist, B., Gerber, S.A., Turecek, F., Gelb, M.H., and Aebersold, R. *Nat. Biotech.* **1999,** *17*, 994.
4. Conrads, T. P., Alving, K., Veenstra, T.D. Belov, M.E., Anderson, G.A., Anderson, D.J., Lipton, M.S., Pasa-Tolic, L., Udseth, H.R., Chrisler, W.B., Thrall, B.D., and Smith, R.D. *Anal. Chem.* **2001,** *73*, 2132.
5. Yao, X. D., Freas, A., Ramirez, J., Demirev, P.A., and Fenselau, C. *Anal. Chem.* **2001,** *73*, 2836.
6. Shen, Y. F., Zhao, R., Belov, M.E., Conrads, T.P., Anderson, G.A., Tang, K.Q., Pasa-Tolic, L. Veenstra, T.D., Lipton, M.S., Udseth, H.R., Smith, R.D. *Anal. Chem.* **2001,** *73*, 1766.
7. Smith, R. D., Pasa-Tolic, L., Lipton, M.S., Jensen, P.K., Anderson, G.A., Shen, Y.F., Conrads, T.P., Udseth, H.R., Harkewicz, R., Belov, M.E., Masselson, C., Veenstra, T.D. *Electrophoresis* **2001,** *22*, 1652.
8. Mortz, E., O'Connore, P.B., Roepstroff, P.Kelleher, N.L., Wood, T.D., McLafferty, F.W., and Mann, M. *Proc. Natl. Acad. Sci. USA* **1996,** *93*, 8264-8267.

APPENDIX A

Matched Peptide Mass (Da)	Error (mmu)	Amino Acid Sequence	Protein	Protein Mass (Da)	Missed Cleavages
967.4709		Hits within 5 ppm (+/-0.0048)			
967.4671	3.79	SIDFQVCR	gene931	26737.3798	0
967.4671	3.79	MVYGENVR	gene111	28315.0767	0
967.4709		Hits within 10 ppm (+/-0.0097)			
**************					
967.4730		Hits within 5 ppm (+/-0.0048)			
967.4730		Hits within 10 ppm (+/-0.0097)			
967.4671	5.89	SIDFQVCR	gene931	26737.3798	0
967.4671	5.89	MVYGENVR	gene111	28315.0767	0
**************					
969.5319		Hits within 5 ppm (+/-0.0048)			
969.5304	1.53	KHAIELCR	gene791	17447.5153	1
969.5319		Hits within 10 ppm (+/-0.0097)			
**************					
971.5263		Hits within 5 ppm (+/-0.0049)			
971.5274	-1.08	EAGAPRDKK	gene183	54564.1695	2
971.5274	-1.09	VTDIRDPR	gene1304	42131.083	1
971.5274	-1.09	VSDRPLER	gene1121	35836.7577	0
971.5274	-1.09	IANQVAEAR	gene459	27234.3934	0
971.5263		Hits within 10 ppm (+/-0.0097)			
971.5202	6.13	EAVEWPIK	gene1778	87184.0306	0
971.5202	6.12	YLGYGTAVK	gene1173	8587.3632	0
971.5202	6.11	SNAYYILK	gene308	24205.6151	0
971.5314	-5.12	HINIPSYK	gene1682	43364.5909	0
971.5348	-8.49	SIHDLMKK	gene949	33771.663	1

```
    971.5348      -8.49        DIPLQVMR        gene666     65347.7101     0

**************


  **971.6519**              Hits within 5 ppm (+/-0.0049)


  **971.6519**              Hits within 10 ppm (+/-0.0097)

*************


  **973.5490**              Hits within 5 ppm (+/-0.0049)
    973.5504      -1.44        LLPKTNCGK       gene1341    50241.7585     1
    973.5504      -1.44        LDKMNPKK        gene266     60617.0067     2


  **973.5490**              Hits within 10 ppm (+/-0.0097)
    973.543        5.96        ANIEDKRK        gene448     22156.17       2
    973.543        5.96        ANIEDKRK        gene441     22054.4649     2
    973.543        5.95        SLGNNILSR       gene1609    85906.7936     0
    973.543        5.95        DTLNQRVK        gene1175    20139.0706     1
    973.543        5.95        ININKSER        gene959     62374.2511     1
    973.5543      -5.28        NTKERAVR        gene860     49693.6401     2
    973.5569      -7.95        VIVLAEDSK       gene1695    17231.9761     0
    973.557       -7.95        SEILDVIGK       gene785     137100.7428    0
    973.557       -7.95        LDIGTVIDK       gene130     42836.9063     0
    973.557       -7.95        DILSELGVK       gene57      19521.9952     0
    973.557       -7.95        DLDEIIKK        gene12      41776.9559     1
    973.557       -7.96        EEIIQITK        gene1020    39414.1504     0
    973.557       -7.96        ITLPSSLDK       gene728     48607.4337     0
    973.5583      -9.3         INSRWLGK        gene174     22626.6609     1

**************


  **973.5633**              Hits within 5 ppm (+/-0.0049)


  **973.5633**              Hits within 10 ppm (+/-0.0097)
    973.5543       9.02        NTKERAVR        gene860     49693.6401     2
    973.5569       6.35        VIVLAEDSK       gene1695    17231.9761     0
    973.557        6.35        SEILDVIGK       gene785     137100.7428    0
    973.557        6.35        LDIGTVIDK       gene130     42836.9063     0
    973.557        6.35        DILSELGVK       gene57      19521.9952     0
    973.557        6.35        DLDEIIKK        gene12      41776.9559     1
```

973.557	6.34	EEIIQITK	gene1020	39414.1504	0
973.557	6.34	ITLPSSLDK	gene728	48607.4337	0
973.5583	5	INSRWLGK	gene174	22626.6609	1
973.5682	-4.89	ELLRDSIK	gene1642	48577.0006	1
973.5682	-4.89	KDLENVKK	gene1002	24548.1702	2
973.5682	-4.89	AESLAKNLK	gene961	30852.2577	1
973.5682	-4.89	VLITAATER	gene925	32070.267	0
973.5682	-4.89	RLSEELVK	gene808	17259.4364	1
973.5682	-4.89	KELEVISR	gene562	119460.1967	1
973.5682	-4.9	NIQKITEK	gene1663	48166.5976	1
973.5682	-4.9	LNTTLNGLK	gene1117	8893.9848	0
973.5682	-4.9	TLNEQKIK	gene1083	38219.9068	1
973.5722	-8.91	IIPKYDPK	gene786	46939.9641	1
973.5722	-8.91	DLQPFILK	gene1577	15849.3415	0
973.5722	-8.91	YVPELPKK	gene897	26590.224	1

****

974.5391		Hits within 5 ppm (+/-0.0049)			
974.5344	4.65	ILEEMIAR	gene959	62374.2511	0
974.5344	4.65	LGNLDMAIK	gene810	48246.2438	0
974.5344	4.65	LMAAINDVK	gene535	54977.2428	0
974.5345	4.64	SPALMNLTK	gene926	25881.81	0
974.5383	0.81	TGRKLDER	gene1854	32762.2277	2
974.5383	0.81	AEDSIKRR	gene347	74355.0988	2
974.5383	0.81	RAEDSIKR	gene347	74355.0988	2
974.5383	0.81	ANRGLTSAGK	gene166	22206.9731	1
974.5383	0.81	TGRLISGDR	gene46	41331.6899	1
974.5383	0.8	KVKNTNDR	gene1494	71169.1572	2
974.5383	0.8	RATDQQKK	gene989	29876.7075	2
974.541	-1.87	TEIEEILK	gene535	54977.2428	0
974.5423	-3.21	FRLAENPK	gene561	46208.3092	1
974.5423	-3.22	TLNHSKFK	gene813	23373.7827	1

974.5391		Hits within 10 ppm (+/-0.0097)			
974.5311	8.03	FEELIPAR	gene1786	27631.3657	0
974.5457	-6.59	TMARGITPK	gene886	18513.0996	1

****

APPENDIX B

Matched Peptide Mass (Da)	Error (mmu)	Missed Cleavages	Amino Acid Sequence
gene436	*6218*		
Coverage 47/51	*(%) =*	*92*	
2717.3761	10.2	2	MDTYSYWKYGLLNLQVEVNPKR
1240.7238	-1.5	2	RRLVNQEGLR
1084.6227	-3.39	1	RLVNQEGLR
928.5216	2.72	0	LVNQEGLR
1902.9547	-10.74	1	EVLDELKTIVEENEDK
***********			
gene1842	*16053*		
Coverage 100/140	*(%) =*	*71*	
1352.6632	0.46	1	MSEERIQFNAK
720.4044	0.17	0	IQFNAK
810.4514	-0.38	1	WYVSKK
1631.824	13.18	1	IKIDENTSNEEIAR
1302.752	10.68	0	VLASIEETLSIK
1382.703	12.94	1	IKDFLPFDMEK
1289.6952	2.97	1	VKEEDISGALTK
1062.5319	-4.16	0	EEDISGALTK
1561.8437	12.71	2	SPGTTKKLGTIDDTK
1085.657	9	0	LLTEIVLER
1053.5654	7.28	1	MIEKYIEK
1268.6924	-3.01	2	MIEKYIEKSK
***********			
gene644	*11390*		
Coverage 72/104	*(%) =*	*69*	
1373.6597	11.88	1	MKEGIFFGEGMK
1114.5243	8.23	0	EGIFFGEGMK
1772.9645	17.59	0	LIAIGIVAATGDETATEK
1193.5771	-9.18	1	QMRSGMNELK
2046.0905	-1.97	2	SGMNELKITKEEIVDALR
1286.7319	9.45	1	ITKEEIVDALR
1363.7659	12.1	0	VVLLTSGMPAFTK
875.5388	0.39	2	AMKVVSKL
***********			
gene1532	*8345*		
Coverage 53/77	*(%) =*	*68*	
1289.7073	-9.14	1	MAAIEVGRVCIK
1716.9616	-11.55	2	MAAIEVGRVCIKTLGR
2975.5875	13.29	2	TLGREAGNTCVIVEVLDKNFVVIDGSVK

1077.5944	5.1	0	NFVVIDGSVK
1537.8225	2.75	2	HVEPTDKKVDLEK
731.4303	1.6	1	KVDLEK
***********			
gene987	*10873*		
Coverage 63/93	*(%) =*	*67*	
1287.6342	-7	2	KINRYCPYCK
1078.5757	-1.44	1	KHTAHTVER
1277.7078	9.99	2	KHTAHTVERAK
1505.7977	9.11	2	ASELKWGQRQFR
2283.3089	-20.31	2	RVTAGYGGYPRPLPGGAKPIKK
864.444	0.8	1	SNGGFRAR
***********			
gene801	*6097*		
Coverage 36/53	*(%) =*	*67*	
868.4528	-8.34	1	TKYGQGSK
1559.8844	-13.09	2	GPGIIRKYGLNLCR
838.4245	4.38	0	YGLNLCR
1091.5672	-8.35	1	QCFRELAPK
1002.5988	0.95	1	ELAPKLGFK
1130.6937	5.29	2	ELAPKLGFKK
***********			
gene353	*12249*		
Coverage 73/109	*(%) =*	*66*	
1134.6159	9.53	0	LFVESQNLGK
1639.8695	15.7	0	AINALSEGGISGFYLK
1355.5578	10.53	0	EYQGMSPDDWK
1447.7506	11.47	0	GFLLAEEPEMAIK
2645.3496	14.51	2	KLENDRYTLVELPVLGMEVNSPE
***********			
gene845	16380		
Coverage 95/144	*(%) =*	*65*	
2557.1703	24.02	0	MNEELQNQFMALDVYNQQVDK
993.4927	8.55	0	SIESMEGLK
1341.8145	3.56	0	EILLPLGAGAFIK
1291.6534	-10.27	0	NVDDVIVDFQK
833.4256	1.59	0	SVEELEK
1314.7381	3.38	1	TKELVNNQIQK
1085.5955	2.22	0	ELVNNQIQK
831.4576	-0.09	0	TNQEIVK
1100.6428	-2.56	1	TNQEIVKLR
874.4998	-5.88	1	LRSELEK
1004.5199	-9.08	1	SAQMKPKTN

****

gene1560	7244		
Coverage 41/63	*(%) =*	*65*	
1074.5795	1.62	1	RTGEELIEK
918.4784	0.43	0	TGEELIEK
1244.5799	9.22	1	FTSDFEENKK
1671.9467	-10.31	1	KAVEEVAMISTKPLR
1543.8517	13.95	0	AVEEVAMISTKPLR
822.4725	-0.3	0	IAGYVTAK

****

gene1734	9740		
Coverage 63/99	*(%) =*	*63*	
1656.8671	5.03	0	MEYIYAALLLNSAGK
1369.7439	12.3	0	AILVAGGVEANEAR
1711.9481	3.58	0	ALVAALEGVDIAEAIEK
1519.7756	9.12	1	KEDTGAAAAAGLGALFG

****

gene36	12488		
Coverage 71/112	*(%) =*	*63*	
1258.7271	6.56	1	KIEAIFRPER
1130.6322	6.92	0	IEAIFRPER
1141.6441	9.46	1	GRGVQGGIIER
928.5216	2.73	0	GVQGGIIER
1579.8484	15.72	1	YRGNEYLVDLLPK
1485.7912	8.65	0	LISENAVTGKPGDGK
1299.7676	7.8	0	IFVIPVEDVIR
1217.649	1.84	2	IRTNETGKDAI

****

gene884	9623		
Coverage 53/86	*(%) =*	*61*	
938.4657	4.1	0	MDAFDVIK
2053.0713	-12.92	2	MDAFDVIKTPIVSEKTMK
1262.6414	-3.84	1	TMKLIEEENR
1937.0496	4.91	2	LIEEENRLVFYVERK
925.5147	4.5	0	LVFYVER
2261.1341	0.9	1	AYITLKDEYNAGEVAASLGIY

****